

THE COMMUNICATIONS HANDBOOK

Editor-in-Chief

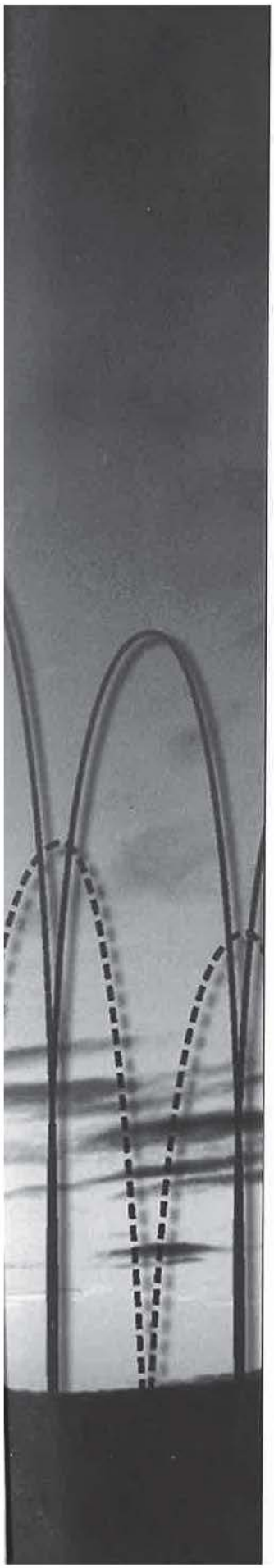
Jerry D. Gibson

CRC PRESS

IEEE PRESS

A CRC Handbook Published in Cooperation with IEEE Press

Samsung Exhibit 1022, Page 1 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540



THE COMMUNICATIONS HANDBOOK

Samsung Exhibit 1022, Page 2 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The Electrical Engineering Handbook Series

Series Editor

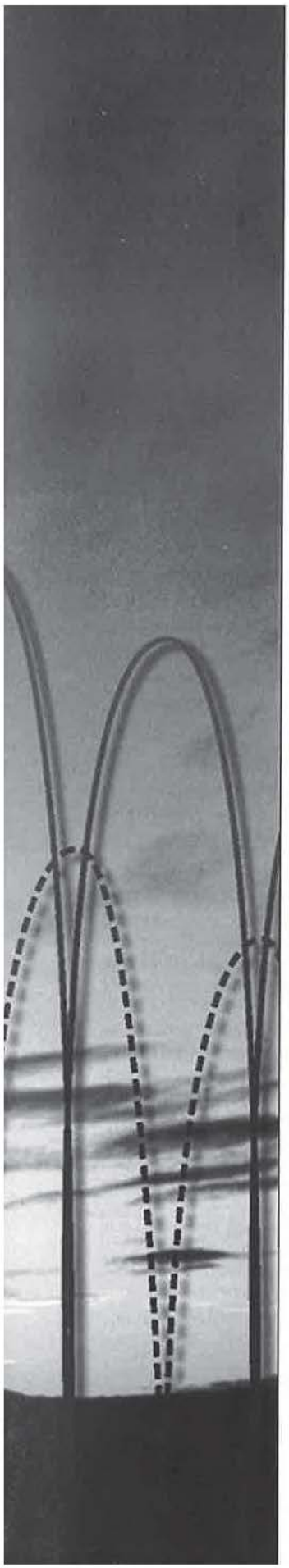
Richard C. Dorf

University of California, Davis

Titles Included in the Series

The Electrical Engineering Handbook, Richard C. Dorf
The Biomedical Engineering Handbook, Joseph D. Bronzino
The Circuits and Filters Handbook, Wai-Kai Chen
The Transforms and Applications Handbook, Alexander D. Poularikas
The Control Handbook, William S. Levine
The Electronics Handbook, Jerry C. Whitaker
The Industrial Electronics Handbook, J. David Irwin
The Communications Handbook, Jerry D. Gibson
The Mobile Communications Handbook, Jerry D. Gibson

Samsung Exhibit 1022, Page 3 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540



THE COMMUNICATIONS HANDBOOK

Editor-in-Chief

Jerry D. Gibson

CRC PRESS

IEEE PRESS

A CRC Handbook Published in Cooperation with IEEE Press

Samsung Exhibit 1022, Page 4 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Library of Congress Cataloging-in-Publication Data

The communications handbook / editor-in-chief, Jerry D. Gibson.
p. cm. -- (Electrical engineering handbook series)
Includes bibliographical references and index.
ISBN 0-8493-8349-8 (alk. paper)
1. Telecommunication--Handbooks, manuals, etc. I. Gibson, Jerry D. II. Series.
TK5101.C6583 1996
621.382--dc20 96-27623
CIP

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without prior permission in writing from the publisher.

All rights reserved. Authorization to photocopy items for internal or personal use, or the personal or internal use of specific clients, may be granted by CRC Press LLC, provided that $.50 per page photocopied is paid directly to Copyright Clearance Center, 27 Congress Street, Salem, MA 01970 USA. The fee code for users of the Transactional Reporting Service is ISBN 0-8493-8349-8/97/$0.00+$.50. The fee is subject to change without notice. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

The consent of CRC Press does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from CRC Press for such copying.

Direct all inquiries to CRC Press LLC, 2000 Corporate Blvd., N.W., Boca Raton, Florida 33431.

© 1997 by CRC Press LLC

No claim to original U.S. Government works
International Standard Book Number 0-8493-8349-8
Library of Congress Card Number 96-27623
Printed in the United States of America 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Preface

The Handbook series published by CRC Press represents a truly unique approach to disseminating technical information. Starting with the first edition of *The Electrical Engineering Handbook* edited by Richard Dorf and published in 1993, this series is dedicated to the idea that a reader should be able to pull one of these handbooks off the shelf and at least 80% of the time find what they need to know about a subject area. As handbooks, these books are also different in that they are more than just a dry listing of facts and data, filled mostly with tables. In fact, a hallmark of these handbooks is that the articles or chapters are designed to be relatively short, written as tutorials or overviews, so that once the reader locates the broader topic, it is easy to find an answer to a specific question.

At the inception of this project, I prepared an outline containing the main headings and chapter titles, about 135 chapters. With the counsel of the Advisory Board, this first table of contents was refined and I began to seek authors. My approach here was fairly simple—ask the best people I know. I then had the opportunity to talk to these very active people about their field. Sometimes as we spoke, prospective authors would indicate that a particular chapter topic or two included in the original outline could be eliminated or combined with another, or was being superceded by later developments. In response, the contents were modified accordingly.

Of course, the authors are the key to achieving the overall goal of the handbook, and having read all of the chapters personally, the results are impressive. The chapters are authoritative, to-the-point, and enjoyable to read. Answers to frequently asked questions, facts, and figures are available almost at a glance. Since the authors are experts in their field, it is understandable that the content is excellent. Additionally, the authors were encouraged to put some of their own interpretations and insights into the chapters, which greatly enhances the readability. However, I am most impressed by the ability of the authors to condense so much information into so few pages. These chapters are unlike any most people have ever written—they are not research journal articles, they are not textbooks, they are not long tutorial review articles for magazines. They really are a new format.

In reading drafts of the chapters, I applied two tests. If the chapter covered a topic that I felt familiar with, I checked to see if it contained what I thought were the essential facts and ideas. If the chapter was in an area of communications less familiar to me, I looked for definitions of terms that I had heard of, or for a discussion that brought forth to me why this area is important and what is happening in the field today. I was amazed at what I learned. You will be too.

Using *The Communications Handbook* is simple. Look up the topic of interest to you either in the Table of Contents or in the Index. Go directly to that chapter. As you are reading the chosen chapter, you may wish to refer to other chapters in the same main heading. If you need some background information in the general communications field, you need only consult the Basic Principles section. There is no need to read the *Handbook* beginning-to-end, start-to-finish. Look up what you need right now, read it, and go back to the task at hand.

Samsung Exhibit 1022, Page 6 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The pleasure of such a project as this is in working with the authors, and I am gratified to have had this opportunity and to be associated with each of them. Additionally, Mary Kugler at CRC Press and the personnel at TechBooks have been extraordinarily helpful and patient as the handbook evolved and was finally brought to fruition. Elaine M. Gibson served as Managing Editor of the *Handbook*, working directly with the authors, the publisher, and the production staff, using her creative talents and enthusiasm to improve operations, solve problems, and make each step of the project as seamless as possible for everyone involved. Elaine and I also acknowledge our editors at CRC Press, first Joel Claypool, and now Kristen Maus, for their insights, guidance, and encouragement in making *The Communications Handbook* a reality.

Jerry D. Gibson
College Station, Texas

Samsung Exhibit 1022, Page 7 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Editor-in-Chief

Jerry D. Gibson currently holds the J.W. Runyon, Jr. Professorship in the Department of Electrical Engineering at Texas A&M University. He has held positions at General Dynamics—Fort Worth (1969–72), the University of Notre Dame (1973–74), and the University of Nebraska—Lincoln (1974–76), and during the fall of 1991, Dr. Gibson was on sabbatical with the Information Systems Laboratory and the Telecommunications Program in the Department of Electrical Engineering at Stanford University.

He is coauthor of the book *Introduction to Nonparametric Detection with Applications* (Academic Press, 1975 and IEEE Press, 1995) and author of the textbook, *Principles of Digital and Analog Communications* (Prentice-Hall, second ed., 1993). He was Associate Editor for Speech Processing for the *IEEE Transactions on Communications* from 1981–1985 and Associate Editor for Communications for the *IEEE Transactions on Information Theory* from 1988–1991. He has served as a member of the Speech Technical Committee of the IEEE Signal Processing Society (1992–1995) and is currently a member of the IEEE Information Theory Society Board of Governors (1990–1996) and a member of the Editorial Board for the *Proceedings of the IEEE*. He is serving as President of the IEEE Information Theory Society in 1996.

Dr. Gibson is Editor-in-Chief of *The Mobile Communications Handbook* (CRC Press, 1995), Editor-in-Chief of *The Communications Handbook* (CRC Press, 1996), and Editor of the IEEE Press book series on Signal Processing.

In 1990, Dr. Gibson received The Frederick Emmons Terman Award from The American Society for Engineering Education, and in 1992, was elected Fellow of the IEEE "for contributions to the theory and practice of adaptive prediction and speech waveform coding." He was co-recipient of the 1993 IEEE Signal Processing Society Senior Paper Award for the Speech Processing area.

His research interests include data, speech, image, and video compression, multimedia over networks, wireless communications, information theory, and digital signal processing.

Samsung Exhibit 1022, Page 8 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Advisory Board

Vijay K. Bhargava received the B.Sc., M.Sc., and Ph.D. degrees from Queen's University of Kingston, Canada in 1970, 1972 and 1974 respectively. He is a Professor of Electrical and Computer Engineering at the University of Victoria, co-author of the book *Digital Communications by Satellite* (New York: Wiley, 1981), and co-editor of the IEEE Press Book *Reed-Solomon Codes and Their Applications*. Dr. Bhargava is the founding President of Binary Communications, Inc. (est. 1983), a company which has successfully produced a programmable (31,k) Reed-Solomon CODEC using VLSI technology. He is very active in the IEEE and is currently a Director of the IEEE and a member of the Board of Governors of the IEEE Information Theory Society.

Chris D. Heegard received the B.S. and the M.S. degrees in electrical and computer engineering from the University of Massachusetts, Amherst, MA, in 1975 and 1976, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1981.

From 1976–1978, he was an R&D Engineer at Linkabit Corp., San Diego, CA, and in 1981, he joined Cornell University, Ithaca, NY, as an Assistant Professor; he was appointed to Associate Professor in 1987.

In 1984 Dr. Heegard received the Presidential Young Investigator Award from the National Science Foundation and the IBM Faculty Development Award. In 1986, he was elected to the Board of Governors of the Information Theory Society of the IEEE, reelected in 1989, and he served as President in 1994.

Laurence B. Milstein received the B.E.E. degree from the City College of New York, New York, NY, in 1964, and the M.S. and Ph.D. degrees in electrical engineering from the Polytechnic Institute of Brooklyn, Brooklyn, NY, in 1966 and 1968, respectively.

From 1968 to 1974, he was employed by the Space and Communications Group of Hughes Aircraft Company, and from 1974 to 1976 he was a member of the Department of Electrical and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY. Since 1976, he has been with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, where he is a Professor and former Department Chairman. He was an Associate Editor for Communication Theory for the *IEEE Transactions on Communications* and an Associate Technical Editor for the *IEEE Communications Magazine*, and he is currently a Senior Editor for the *IEEE Journal on Selected Areas in Communications*. He was the Vice-President for Technical Affairs in 1990 and 1991 of the IEEE Communications Society, and has served as a member of the Board of Governors of both the IEEE Communications Society and the IEEE Information Theory Society.

James W. Modestino received the B.S. degree from Northeastern University, Boston, MA, in 1962, and the M.S. degree from the University of Pennsylvania, PA in 1964, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 1968 and 1969, respectively.

Samsung Exhibit 1022, Page 9 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

He has held industrial positions with RCA Communication Systems Division, Camden, NJ; AVCO Systems Division, Wilmington, MA; GTE Laboratories, Inc., Waltham, MA and M.I.T. Lincoln Laboratory, Lexington, MA. From 1970 to 1972, he was an Assistant Professor in the Department of Electrical Engineering, Northeastern University. In 1972, he joined Rensselaer Polytechnic Institute, Troy, NY, where he is currently a Professor in the Department of Electrical, Computer, and Systems Engineering. Dr. Modestino is a past Associate Editor of the *IEEE Transactions on Information Theory*, and in 1984, he was co-recipient of the Stephen O. Rice Prize Paper Award from the IEEE Communications Society.

Joseph C. Palais has been a member of the electrical engineering faculty at Arizona State University since 1964, where is presently serving as a professor. He received the B.S.E.E. degree from the University of Arizona, Tucson, in 1959 and the M.S.E. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1962 and 1964 respectively.

Professor Palais has worked as a Microwave Engineer for Motorola and Stanford Research Institute and has been a consultant to several firms, including Sperry Flight Systems, Sylvania, and Polaroid. In 1967 he was awarded a Ford Foundation Residency in Engineering Practice, and in 1973 he served as a Visiting Associate Professor at the Technion-Israel Institute of Technology.

In 1993, he received the IEEE Educational Activities Board Meritorious Achievement Award in Continuing Education.

Dr. Palais is the author of the textbook *Fiber Optic Communications*, Prentice-Hall, 3rd edition 1992. Professor Palais has also coauthored the software manual, *Fiber Optic Systems and Devices with Microcomputer Software*, with D. Johnson, Kern International, Inc., 1987.

Jonathan S. Turner received the M.S. and Ph.D. degrees in computer science from Northwestern University in 1979 and 1981.

He is a Professor and Chairman of the Department of Computer Science at Washington University, where he has been since 1983. His primary research interest is the design and analysis of switching systems, with special interest in systems supporting multicast communication. He has been awarded more than a dozen patents for his work on switching systems and has a number of widely cited publications in this area. His research interests also include the study of algorithms and computational complexity, with particular interest in the probably performance of heuristic algorithms for NP-complete problems.

Dr. Turner is a member of the ACM and SIAM.

Managing Editor
Elaine M. Gibson

Samsung Exhibit 1022, Page 10 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Contributors

Saf Asghar
Advanced Micro Devices, Inc.
Austin, Texas

Chris B. Autry
Georgia Institute of Technology
Atlanta, Georgia

Joseph Bannister
The Aerospace Corporation
Los Angeles, California

Melbourne Barton
Bell Communications Research
Morristown, New Jersey

Vijay K. Bhargava
University of Victoria
Victoria, B.C., Canada

Ezio Biglieri
Politecnico di Torino
Torino, Italia

Anders Bjarklev
Technical University of Denmark
Lyngby, Denmark

Daniel J. Blumenthal
Georgia Institute of Technology
Atlanta, Georgia

Paula M. Bordogna
Lucent Technologies
North Andover, Massachusetts

Madhukar Budagavi
Texas A&M University
College Station, Texas

Pierre Catala
Texas A&M University
College Station, Texas

Wai-Yip Chan
Illinois Institute of Technology
Chicago, Illinois

Biao Chen
University of Texas-Dallas
Dallas, Texas

Giovanni Cherubini
IBM Zurich Research Laboratory
Ruschlikon, Switzerland

Stanley Chia
BT Laboratories
Ipswich, United Kingdom

Youn Ho Choung
TRW Space & Defense
Redondo Beach, California

John M. Cioffi
Stanford University
Stanford, California

Leon W. Couch
University of Florida
Gainesville, Florida

Rene L. Cruz
University of California-San Diego
La Jolla, California

Marc Delprat
Alcatel Telecom
Mobile Communication Division
Colombes, France

Paul Diament
Columbia University
New York, New York

Spiros Dimolitsas
Lawrence Livermore National Laboratory
Livermore, California

Robert L. Douglas
University of Memphis
Memphis, Tennessee

Eric Dubois
University of Quebec
Verdun, Quebec, Canada

Niloy K. Dutta
Lucent Technologies
Murray Hill, New Jersey

Bruce Elbert
Hughes Space and Communications International, Inc.
El Segundo, California

Ahmed K. Elhakeem
Concordia University
Montreal, Quebec, Canada

Ivan J. Fair
Technical University of Nova Scotia
Halifax, Nova Scotia, Canada

John L. Fike
Texas A&M University
College Station, Texas

Michael D. Floyd
Motorola Semiconductor Products
Austin, Texas

Lew E. Franks
University of Massachusetts
Amherst, Massachusetts

Susan A.R. Garrod
Purdue University
West Lafayette, Indiana

Samsung Exhibit 1022, Page 11 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Costas Georghiades
Texas A&M University
College Station, Texas

Ira A. Gerson
Motorola Semiconductor Products
Schaumburg, Illinois

Klein Gilhousen
Qualcomm, Inc.
San Diego, California

David Haccoun
Ecole Polytechnique de Montreal
Montreal, Quebec, Canada

Frederick Halsall
University of Wales
Swansea, Wales, United Kingdom

Jeff Hamilton
General Instrument Corporation
Hatboro, Pennsylvania

Lajos Hanzo
University of Southampton
Highfield, Southampton
United Kingdom

Roger Haskin
IBM Almaden Research Center
San Jose, California

Tor Helleseth
University of Bergen
Bergen, Norway

Garth D. Hillman
Motorola, Inc.
Austin, Texas

Stephen J. Hinterlong
Lucent Technologies
Naperville, Illinois

Michael L. Honig
Northwestern University
Evanston, Illinois

Hwei P. Hsu
Farleigh Dickinson University
Teaneck, New Jersey

Erwin Hudson
TRW Electronics Systems & Technology Division
Redondo Beach, California

Yeongming Hwang
City University of Hong Kong
Kowloon, Hong Kong

Louis J. Ippolito, Jr.
Stanford Telecom Inc.
Reston, Virginia

Bijan Jabbari
George Mason University
Fairfax, Virginia

Ravi Jain
Bell Communications Research
Red Bank, New Jersey

Amos E. Joel, Jr.
Executive Consultant
South Orange, New Jersey

Varun Kapoor
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

B.L. Kasper
Lucent Technologies
Breiningsville, Pennsylvania

Kota Kinoshita
NTT Do Co Mo
Yokosuka Shi, Japan

Mark Kolber
General Instrument Corporation
Hatboro, Pennsylvania

Boneung Koo
Kyonggi University
Kyonggi-Do, Korea

Nitin C. Kothari
Georgia Institute of Technology
Atlanta, Georgia

Andrew D. Kucar
4U Comm. Research Inc.
Ottawa, Ontario, Canada

Vijay P. Kumar
University of Southern California
Los Angeles, California

Vinod Kumar
Alcatel Telecom
Mobile Communication Division
Colombes, France

B.P. Lathi
California State University
Sacramento, California

Allen H. Levesque
GTE Laboratories, Inc.
Needham, Massachusetts

Curtis A. Levis
Ohio State University
Columbus, Ohio

Chung Sheng Li
IBM T.J. Watson Research Center
Yorktown Heights, New Jersey

Yi-Bing Lin
National Chaio Tung University
Hsinchu, Taiwan, R.O.C.

Joseph L. LoCicero
Illinois Institute of Technology
Chicago, Illinois

John H. Lodge
Communications Research Centre
Ottawa, Ontario, Canada

Mari Maeda
Bell Communications Research
Red Bank, New Jersey

Nicholas Malcolm
Hewlett-Packard (Canada) Ltd.
Burnaby, BC, Canada

Masud Mansuripur
University of Arizona
Tucson, Arizona

Gerald A. Marin
IBM
Triangle Park, North Carolina

Nasir D. Memon
Northern Illinois University
DeKalb, Illinois

Paul Mermelstein
Bell Northern Research
Verdun, Quebec, Canada

Toshio Miki
Mobile Communication Network, Inc.
Yokosuka-shi, Kanagawa, Japan

Laurence B. Milstein
University of California at San Diego
La Jolla, California

Abdi R. Modarressi
Bell Laboratories
Columbus, Ohio

Seshadri Mohan
Bell Communications Research
Morristown, New Jersey

Samsung Exhibit 1022, Page 12 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Michael Moher
Communications Research Centre
Ottawa, Ontario, Canada

Jaekyun Moon
University of Minnesota
Minneapolis, Minnesota

Rias Muhamed
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Masao Nakagawa
Keio University
Hiyoshi, Yokohama, Japan

Madihally J. Narasimha
Telecom Solutions
San Jose, California

A. Michael Noll
University of Southern California
Los Angeles, California

Peter Noll
Technische Universitaet Berlin,
Berlin, West Germany

Peter P. Nuspl
W.L. Pritchard & Co.
Bethesda, Maryland

Michael O'Flynn
San Jose State University
San Jose, California

Raif Onvural
Duke University
Durham, North Carolina

Geoffrey C. Orsak
George Mason University
Fairfax, Virginia

Henry W.L. Owen III
Georgia Institute of Technology
Atlanta, Georgia

Kaveh Pahlavan
Worcester Polytechnic Institute
Worcester, Massachusetts

Joseph C. Palais
Arizona State University
Tempe, Arizona

Bernd-Peter Paris
George Mason University
Fairfax, Virginia

Lance Parr
Texas A&M University
College Station, Texas

Bhasker P. Patel
Illinois Institute of Technology
Chicago, Illinois

Achille Pattavina
Politecnico di Milano
Milano, Italy

Arogyaswami Paulraj
Stanford University
Stanford, California

Ken D. Pedrotti
Rockwell International
Thousand Oaks, California

Eric Petajan
Lucent Technologies
Murray Hill, New Jersey

Roman Pichna
University of Victoria
Victoria, B.C., Canada

Samuel Pierre
Universite du Quebec
Montreal, Quebec, Canada

Alistair J. Price
Rockwell International
Thousand Oaks, California

David R. Pritchard
Norlight Telecommunications/ Teleport Chicago
Skokie, Illinois

Wilbur L. Pritchard
W.L. Pritchard & Co.
Bethesda, Maryland

John G. Proakis
Northeastern University
Boston, Massachusetts

Ramesh R. Rao
University of California at San Diego
La Jolla, California

Theodore S. Rappaport
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Erwin P. Rathgeb
Siemens AG
Muenchen, Germany

Narasimha Reddy
Texas A&M University
College Station, Texas

Whitham D. Reeve
Reeve Engineers
Anchorage, Alaska

Bixio Rimoldi
Washington University
St. Louis, Missouri

Thomas G. Robertazzi
State University of New York
Stony Brook, New York

Martin S. Roden
California State University
Los Angeles, California

Arthur H.M. Ross
Qualcomm, Inc.
San Diego, California

Izhak Rubin
University of California at Los Angeles
Los Angeles, California

Khalid Sayood
University of Nebraska
Lincoln, Nebraska

Charles Schell
General Instrument Corporation
Hatboro, Pennsylvania

A. Udaya Shankar
University of Maryland
College Park, Maryland

Curtis A. Siller, Jr.
Lucent Technologies
North Andover, Massachusetts

Marvin K. Simon
Jet Propulsion Laboratory
Pasadena, California

Samsung Exhibit 1022, Page 13 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Suresh P. Singh
University of South Carolina
Columbia, South Carolina

Bernard Sklar
Communications Engineering Services
Tarzana, California

R.A. Skoog
AT&T
Holmdel, New Jersey

David R. Smith
George Washington University
Ashburn, Virginia

Jonathan M. Smith
University of Pennsylvania
Philadelphia, Pennsylvania

Richard G. Smith
Center Valley, Pennsylvania

Raymond Steele
Multiple Access Communications Limited
Southampton, Hampshire
United Kingdom

Ferrel G. Stremler
University of Wisconsin
Madison, Wisconsin

Gordon Stuber
Georgia Institute of Technology
Atlanta, Georgia

Len Taupier
General Instrument Corporation
Hatboro, Pennsylvania

Chong Kwan Un
Korea Advanced Institute of Science and Technology
Taejon, Korea

Reudiger Urbanke
Lucent Technologies
Murray Hill, New Jersey

Harrell J. Van Norman
Dayton, Ohio

David Vlack
Bell Labs
St. Charles, Illinois

Qiang Wang
University of Victoria
Victoria, B.C., Canada

Richard H. Williams
University of New Mexico
Albuquerque, New Mexico

Ping Wah Wong
Hewlett Packard Laboratories
Palo Alto, California

Maynard A. Wright
Electrical Engineering Consultant
Citrus Heights, California

Michel D. Yacoub
The University of Campinas
San Paulo, Brazil

Shinj Yamashita
RCAST University of Tokyo
Tokyo, Japan

Chong Ho Yoon
University of Arizona
Tucson, Arizona

William C. Young
Bell Communications Research
Red Bank, New Jersey

Wei Zhao
Texas A&M University
College Station, Texas

Rodger E. Ziemer
University of Colorado at Colorado Springs
Colorado Springs, Colorado

Samsung Exhibit 1022, Page 14 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Contents

SECTION I Basic Principles

1 Analog Modulation *Ferrel G. Stremler* 3

2 Sampling *Hwei P. Hsu* 13

3 Pulse Code Modulation *Leon W. Couch II* 23

4 Probabilities and Random Variables *Michael O'Flynn* 35

5 Random Processes, Autocorrelation, and Spectral Densities *Lew E. Franks* 62

6 Queuing *Richard H. Williams* 78

7 Multiplexing *Martin S. Roden* 87

8 Pseudonoise Sequences *Tor Helleseth and P. Vijay Kumar* 94

9 D/A and A/D Converters *Susan A.R. Garrod* 107

10 Signal Space *Rodger E. Ziemer* 118

11 Channel Models *David R. Smith* 131

12 Optimum Receivers *Geoffrey C. Orsak* 141

13 Standards Setting Bodies *Spiros Dimolitsas* 155

14 Forward Error Correction Coding *Vijay Bhargava and Ivan J. Fair* 166

15 Automatic Repeat Request *David Haccoun and Samuel Pierre* 181

16 Spread Spectrum Communications *Laurence B. Milstein and Marvin K. Simon* 199

17 Diversity Techniques *Arogyaswami Paulraj* 213

18 Information Theory *Bixio Rimoldi and Reudiger Urbanke* 224

Samsung Exhibit 1022, Page 15 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

19 Digital Communication System Performance *Bernard Sklar* 236
20 Synchronization *Costas N. Georghiades* 255
21 Digital Modulation Techniques *Ezio Biglieri* 273

SECTION II Telephony

22 POTS (Plain Old Telephone Service) *A. Michael Noll* 291
23 FDM Hierarchy *Pierre Catala* 301
24 Analog Telephone Channels and the Subscriber Loop *Whitham D. Reeve* 308
25 Baseband Signalling and Pulse Shaping *Michael L. Honig and Melbourne Barton* 318
26 Channel Equalization *John G. Proakis* 339
27 PCM Codec Filters *Michael D. Floyd and Garth D. Hillman* 364
28 Digital Hierarchy *B.P. Lathi and Maynard Wright* 377
29 Line Coding *Joseph L. LoCicero and Bhasker P. Patel* 386
30 Telecommunications Network Synchronization *Madihally J. Narasimha* ... 404
31 Echo Cancellation *Giovanni Cherubini* 414
32 Switching Fabrics *Amos E. Joel, Jr.* 425
33 Customer Premises Equipment *John L. Fike and M. Lance Parr* 433
34 Asymmetric Digital Subscriber Lines *John M. Cioffi* 450
35 Overview of Common Channel Signalling *Abdi R. Modarressi and R.A. Skoog* 480
36 Digital Cross-Connect Systems *Paula M. Bordogna and Curtis A. Siller, Jr.* 496
37 Building Future Networks by Using Photonics in Switching *Stephen J. Hinterlong and David Vlack* 513
38 Asynchronous Transfer Mode (ATM) *Raif O. Onvural* 529
39 Synchronous Optical Network (SONET) *Henry L. Owen and Chris B. Autry* 542
40 Synchronous Digital Hierarchy (SDH) *Henry L. Owen and Chris B. Autry* 554

Samsung Exhibit 1022, Page 16 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

SECTION III Networks

41 The OSI Seven-Layer Model *Fred Halsall* 567

42 ISDN and B-ISDN *Erwin P. Rathgeb* 577

43 Ethernet Networks *Ramesh R. Rao* 591

44 FDDI *Biao Chen, Wei Zhao, and Nicholas Malcolm* 597

45 Broadband Local Area Networks *Joseph A. Bannister* 611

46 Multiple Access Methods *Izhak Rubin* 622

47 Routing and Flow Control *Rene L.Cruz* 650

48 Transport Layer *A. Udaya Shankar* 661

49 Gigabit Networks *Jonathan M. Smith* 672

50 Local Area Networks *Thomas G. Robertazzi* 681

51 Asynchronous Time Division Switching *Achille Pattavina* 686

52 Internetworking *Harrell J. Van Norman* 701

53 Architectural Framework for ATM Networks: Broadband Network Services *Gerald A. Marin and Raif O. Onvural* 717

SECTION IV Optical

54 Fiber Optic Communications Systems *Joseph C. Palais* 731

55 Optical Fibers and Lightwave Propagation *Paul Diament* 740

56 Optical Sources *Niley K. Dutta* 751

57 Optical Transmitters *Alistair J. Price and Ken D. Pedrotti* 774

58 Optical Receivers *Richard G. Smith and B.L. Kasper* 789

59 Fiber Optic Connectors and Splices *William C. Young* 803

60 Passive Optical Components *Joseph C. Palais* 824

61 Semiconductor Optical Amplifiers *Daniel J. Blumenthal and Nitin C. Kothari* 832

62 Optical Amplifiers *Anders Bjarklev* 848

63 Coherent Systems *Shinji Yamashita* 862

Samsung Exhibit 1022, Page 17 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

64 Fiber Optic Applications *Chung-Sheng Li* 872

65 Wavelength Division Multiplexed Systems and Applications *Mari W. Maeda* 883

SECTION V Satellite

66 Geostationary Communications Satellites and Applications *Bruce R. Elbert* 893

67 Satellite Systems *Robert L. Douglas* 912

68 The Earth Station *David R. Pritchard* 922

69 Satellite Transmission Impairments *Louis J. Ippolito, Jr.* 935

70 Satellite Link Design *Peter P. Nuspl and Jahangir A. Tehrani* 949

71 The Calculation of System Temperature for a Microwave Receiver *Wilbur L. Pritchard* 966

72 Onboard Switching and Processing *Ahmed K. Elhakeem* 976

73 Path Diversity *Curt A. Levis* 996

74 Mobile Satellite Systems *John H. Lodge and Michael Moher* 1015

75 Satellite Antennas *Yeongming Hwang and Youn Hɔ Choung* 1032

76 Tracking and Data Relay Satellite System (TDRSS) *Erwin C. Hudson* 1054

SECTION VI Wireless

77 Mobile Radio: An Overview *Andrew D. Kucar* 1069

78 Base Station Subsystems *Chong Kwan Un and Chong Ho Yoon* 1090

79 Access Methods *Bernd-Peter Paris* 1104

80 Location Strategies for Personal Communications Services *Ravi Jain, Yi-Bing Lin, and Seshadri Mohan* 1116

81 Cell Design Principles *Michel Daoud Yacoub* 1146

82 Microcellular Radio Communications *Raymond Steele* 1160

83 Fixed and Dynamic Channel Assignment *Bijan Jabbari* 1175

84 Propagation Models *Theodore S. Rappaport, Rias Muhamed, and Varun Kapoor* .. 1182

85 Power Control *Roman Pichna and Qiang Wang* 1197

86 Second Generation Systems *Marc Delprat and Vinod Kumar* 1208

Samsung Exhibit 1022, Page 18 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

87 The Pan-European Cellular System *Lajos Hanzo* 1226
88 The IS-54 Digital Cellular Standard *Paul Mermelstein* 1246
89 CDMA Technology and the IS-95 North American Standard *Arthur H.M. Ross and Klein S. Gilhousen* 1257
90 Japanese Cellular Standard *Kota Kinoshita and Masao Nakagawa* 1276
91 The British Cordless Telephone Standard: CT-2 *Lajos Hanzo* 1289
92 DECT (Digital European Cordless Telephone) *Saf Asghar* 1305
93 The RACE Program *Stanley Chia* 1327
94 Half-Rate Standards *Wai-Yip Chan, Ira Gerson, and Toshio Miki* 1338
95 Modulation Methods *Gordon L. Stüber* 1353
96 Wireless LANs *Suresh P. Singh* 1367
97 Wireless Data *Allen H. Levesque and Kaveh Pahlavan* 1380

SECTION VII Source Compression

98 Lossless Compression *Khalid Sayood and Nasir D. Memon* 1397
99 Facsimile *Nasir D. Memon and Khalid Sayood* 1411
100 Speech *Boneung Koo* 1424
101 Still Image Compression and Halftoning *Ping Wah Wong* 1437
102 Video *Eric Dubois* 1449
103 The High-Definition Television Grand Alliance System *Eric Petajan* 1462
104 Audio Coding *Peter Noll* 1475
105 Cable *Jeff Hamilton, Mark Kolber, Charles Schell, and Len Taupier* 1488
106 Video Servers *Narasimha Reddy and Roger Haskin* 1502
107 Desktop Videoconferencing *Madhukar Budagavi* 1516

SECTION VIII Data Recording

108 Magnetic Storage *Jaekyun Moon* 1531
109 Magneto-Optical Disk Data Storage *Masud Mansuripur* 1546
Index 1567

Samsung Exhibit 1022, Page 19 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

8

Pseudonoise Sequences

Tor Helleseth
University of Bergen

P. Vijay Kumar
University of Southern California

8.1 Introduction 94
8.2 m Sequences 95
8.3 The q-ary Sequences with Low Autocorrelation 97
8.4 Families of Sequences with Low Crosscorrelation 98
Gold and Kasami Sequences • Quaternary Sequences with Low Crosscorrelation • Kerdock Sequences
8.5 Aperiodic Correlation 101
Barker Sequences • Sequences with High Merit Factor • Sequences with Low Aperiodic Crosscorrelation
8.6 Other Correlation Measures 103
Partial-Period Correlation • Mean Square Correlation • Optical Orthogonal Codes

8.1 Introduction

Pseudonoise sequences (PN sequences), also referred to as pseudorandom sequences, are sequences that are deterministically generated and yet possess some properties that one would expect to find in randomly generated sequences. Applications of PN sequences include signal synchronization, navigation, radar ranging, random number generation, spread-spectrum communications, multipath resolution, cryptography, and signal identification in multiple-access communication systems. The *correlation* between two sequences $\{x(t)\}$ and $\{y(t)\}$ is the complex inner product of the first sequence with a shifted version of the second sequence. The correlation is called 1) an autocorrelation if the two sequences are the same, 2) a crosscorrelation if they are distinct, 3) a periodic correlation if the shift is a cyclic shift, 4) an aperiodic correlation if the shift is not cyclic, and 5) a partial-period correlation if the inner product involves only a partial segment of the two sequences. More precise definitions are given subsequently.

Binary ***m* sequences**, defined in the next section, are perhaps the best-known family of PN sequences. The balance, run-distribution, and autocorrelation properties of these sequences mimic those of random sequences. It is perhaps the random-like correlation properties of PN sequences that makes them most attractive in a communications system, and it is common to refer to any collection of low-correlation sequences as a family of PN sequences.

Section 8.2 begins by discussing m sequences. Thereafter, the discussion continues with a description of sequences satisfying various correlation constraints along the lines of the accompanying self-explanatory figure, Fig. 8.1. Expanded tutorial discussions on pseudorandom sequences may be found in Sarwate and Pursley, 1980, in Chapter 5 of Simon et al., 1994, and in Helleseth and Kumar, 1996.

0-8493-8349-8/97/$0.00+$.50
© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 20 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

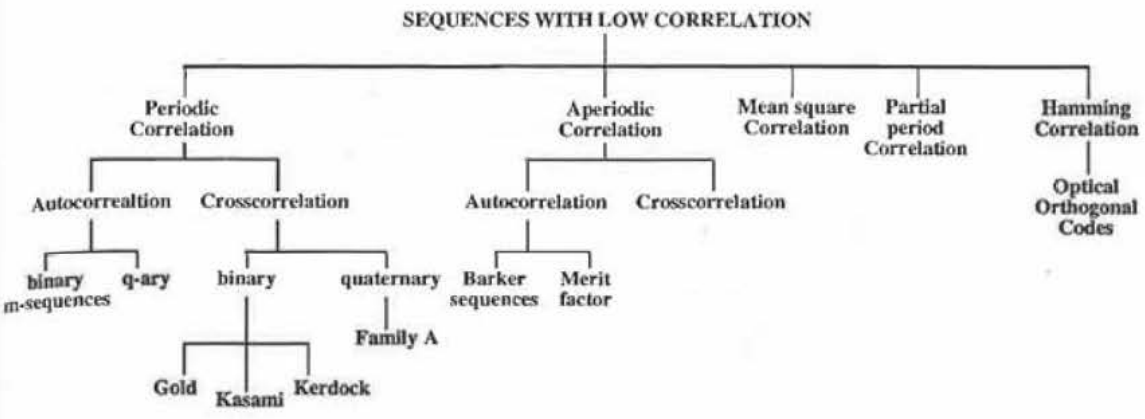


FIGURE 8.1 Overview of pseudonoise sequences.

8.2 *m* Sequences

A binary {0, 1} **shift-register sequence** $\{s(t)\}$ is a sequence that satisfies a linear recurrence relation of the form

$$\sum_{i=0}^{r} f_i s(t+i) = 0, \qquad \text{for all } t \geq 0 \tag{8.1}$$

where $r \geq 1$ is the *degree* of the recursion; the coefficients f_i belong to the finite field $GF(2) = \{0, 1\}$ where the leading coefficient $f_r = 1$. Thus, both sequences $\{a(t)\}$ and $\{b(t)\}$ appearing in Fig. 8.2 are shift-register sequences. A sequence satisfying a recursion of the form in Eq. (8.1) is said to have *characteristic polynomial* $f(x) = \sum_{i=0}^{r} f_i x^i$. Thus, $\{a(t)\}$ and $\{b(t)\}$ have characteristic polynomials given by $f(x) = x^3 + x + 1$ and $f(x) = x^3 + x^2 + 1$, respectively.

Since an r-bit binary shift register can assume a maximum of 2^r different states, it follows that every shift-register sequence $\{s(t)\}$ is eventually periodic with period $n \leq 2^r$, i.e.,

$$s(t) = s(t+n), \qquad \text{for all } t \geq N$$

for some integer N. In fact, the maximum period of a shift-register sequence is $2^r - 1$, since a shift register that enters the all-zero state will remain forever in that state. The upper shift register in Fig. 8.2 when initialized with starting state 0 0 1 generates the periodic sequence $\{a(t)\}$ given by

$$0010111 \quad 0010111 \quad 0010111 \quad \cdots \tag{8.2}$$

of period $n = 7$. It follows then that this shift register generates sequences of maximal period starting from any nonzero initial state.

An m sequence is simply a binary shift-register sequence having maximal period. For every $r \geq 1$, m sequences are known to exist. The periodic **autocorrelation** function θ_s of a binary {0, 1} sequence $\{s(t)\}$ of period n is defined by

$$\theta_s(\tau) = \sum_{t=0}^{n-1} (-1)^{s(t+\tau)-s(t)}, \qquad 0 \leq \tau \leq n-1$$

An m sequence of length $2^r - 1$ has the following attributes. 1) *Balance property:* in each period of the m sequence there are 2^{r-1} ones and $2^{r-1} - 1$ zeros. 2) *Run property:* every nonzero binary s-tuple,

Samsung Exhibit 1022, Page 21 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

An OOC code that achieves the Johnson bound is said to be optimal. A family $\{\mathcal{F}_n\}$ of OOCs indexed by the parameter n and arising from a common construction is said to be asymptotically optimum if

$$\lim_{n\to\infty} \frac{|\mathcal{F}_n|}{M(n, w, \lambda)} = 1$$

Constructions for optical orthogonal codes are available for the cases when $\lambda = 1$ and $\lambda = 2$. For larger values of λ, there exist constructions which are asymptotically optimum. Further details may be found in Helleseth and Kumar, 1996.

Defining Terms

Autocorrelation of a sequence: The complex inner product of the sequence with a shifted version itself.

Crosscorrelation of two sequences: The complex inner product of the first sequence with a shifted version of the second sequence.

***m* Sequence:** A periodic binary {0, 1} sequence that is generated by a shift register with linear feedback and which has maximal possible period given the number of stages in the shift register.

Pseudonoise sequences: Also referred to as pseudorandom sequences (PN), these are sequences that are deterministically generated and yet possess some properties that one would expect to find in randomly generated sequences.

Shift-register sequence: A sequence with symbols drawn from a field, which satisfies a linear-recurrence relation and which can be implemented using a shift register.

References

Baumert, L.D. 1971. *Cyclic Difference Sets*, Lecture Notes in Mathematics 182, Springer–Verlag, New York.

Golomb, S.W. 1982. *Shift Register Sequences*, Aegean Park Press, San Francisco, CA.

Hammons, A.R., Jr. and Kumar, P.V. 1993. On a recent 4-phase sequence design for CDMA. *IEICE Trans. Commun.* E76-B(8).

Helleseth, T. and Kumar, P.V. 1996. (planned). Sequences with low correlation. In *Handbook of Coding Theory*, ed. R. Brualdi, C. Huffman, and V. Pless, Elsevier Science Publishers, Amsterdam.

Jensen, J.M., Jensen, H.E., and Høholdt, T. 1991. The merit factor of binary sequences related to difference sets. *IEEE Trans. Inform. Theory.* IT-37(May):617–626.

MacWilliams, F.J. and Sloane, N.J.A. 1977. *The Theory of Error-Correcting Codes*, North-Holland, Amsterdam.

Mow, W.H. 1994. On McEliece's open problem on minimax aperiodic correlation. In *Proc. IEEE Intern. Symp. Inform. Theory*, p. 75.

Peterson, W.W. and Weldon, E.J., Jr. 1972. *Error-Correcting Codes*, 2nd ed. MIT Press, Cambridge, MA.

Sarwate, D.V. 1984. An upper bound on the aperiodic autocorrelation function for a maximal-length sequence. *IEEE Trans. Inform. Theory.* IT-30(July):685–687.

Sarwate, D.V. and Pursley, M.B. 1980. Crosscorrelation properties of pseudorandom and related sequences. *Proc. IEEE*, 68(May):593–619.

Simon, M.K., Omura, J.K., Scholtz, R.A., and Levitt, B.K. 1994. *Spread Spectrum Communications Handbook*, revised ed. McGraw Hill, New York.

Samsung Exhibit 1022, Page 22 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

9

D/A and A/D Converters

9.1 D/A and A/D Circuits 107
D/A and A/D Converter Performance Criteria • D/A Conversion Processes • D/A Converter ICs • A/D Conversion Processes • A/D Converter ICs • Grounding and Bypassing on D/A and A/D ICs • Selection Criteria for D/A and A/D Converter ICs

Susan A.R. Garrod
Purdue University

Digital-to-analog (D/A) conversion is the process of converting digital codes into a continuous range of analog signals. *Analog-to-digital (A/D) conversion* is the complementary process of converting a continuous range of analog signals into digital codes. Such conversion processes are necessary to interface real-world systems, which typically monitor continuously varying analog signals, with digital systems that process, store, interpret, and manipulate the analog values.

D/A and A/D applications have evolved from predominately military-driven applications to consumer-oriented applications. Up to the mid-1980s, the military applications determined the design of many D/A and A/D devices. The military applications required very high performance coupled with hermetic packaging, radiation hardening, shock and vibration testing, and military specification and record keeping. Cost was of little concern, and "low power" applications required approximately 2.8 W. The major applications up the mid-1980s included military radar warning and guidance systems, digital oscilloscopes, medical imaging, infrared systems, and professional video.

The applications requiring D/A and A/D circuits in the 1990s have different performance criteria from those of earlier years. In particular, low power and high speed applications are driving the development of D/A and A/D circuits, as the devices are used extensively in battery-operated consumer products. The predominant applications include cellular telephones, hand-held camcorders, portable computers, and set-top cable TV boxes. These applications generally have low power and long battery life requirements, or they may have high speed and high resolution requirements, as is the case with the set-top cable TV boxes.

9.1 D/A and A/D Circuits

D/A and A/D conversion circuits are available as integrated circuits (ICs) from many manufacturers. A huge array of ICs exists, consisting of not only the D/A or A/D conversion circuits, but also closely

0-8493-8349-8/97/$0.00+$.50
© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 23 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

TABLE 9.1 D/A and A/D Integrated Circuits

D/A Converter ICs	Resolution, b	Multiplying vs Fixed Reference	Settling Time, μs	Input Data Format
Analog devices AD558	8	Fixed reference	3	Parallel
Analog devices AD7524	8	Multiplying	0.400	Parallel
Analog devices AD390	Quad, 12	Fixed reference	8	Parallel
Analog devices AD1856	16	Fixed reference	1.5	Serial
Burr–Brown DAC729	18	Fixed reference	8	Parallel
DATEL DACHF8	8	Multiplying	0.025	Parallel
National DAC0800	8	Multiplying	0.1	Parallel

A/D Converter ICs	Resolution, b	Signal Inputs	Conversion Speed, μs	Output Data Format
Analog devices AD572	12	1	25	Serial and Parallel
Burr–Brown ADC803	12	1	1.5	Parallel
Burr–Brown ADC701	16	1	1.5	Parallel
National ADC1005B	10	1	50	Parallel
TI, National ADC0808	8	8	100	Parallel
TI, National ADC0834	8	4	32	Serial
TI TLC0820	8	1	1	Parallel
TI TLC1540	10	11	21	Serial

Interface ICs A/D and D/A	Resolution, b	Onboard Filters	Sampling Rate, kHz	Data Format
TI TLC32040	14	Yes	19.2 (programmable)	Serial
TI 2914 PCM codec and filter	8	Yes	8	Serial

related circuits such as sample-and-hold amplifiers, analog multiplexers, voltage-to-frequency and frequency-to-voltage converters, voltage references, calibrators, operation amplifiers, isolation amplifiers, instrumentation amplifiers, active filters, DC-to-DC converters, analog interfaces to digital signal processing systems, and data acquisition subsystems. Data books from the IC manufacturers contain an enormous amount of information about these devices and their applications to assist the design engineer.

The ICs discussed in this chapter will be strictly the D/A and A/D conversion circuits. Table 9.1 lists a small sample of the variety of the D/A and A/D converters currently available. The ICs

Samsung Exhibit 1022, Page 24 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

usually perform either D/A or A/D conversion. There are serial interface ICs, however, typically for high-performance audio and digital signal processing applications, that perform both A/D and D/A processes.

D/A and A/D Converter Performance Criteria

The major factors that determine the quality of performance of D/A and A/D converters are *resolution, sampling rate, speed,* and *linearity.*

The *resolution* of a D/A circuit is the smallest change in the output analog signal. In an A/D system, the resolution is the smallest change in voltage that can be detected by the system and that can produce a change in the digital code. The resolution determines the total number of digital codes, or *quantization levels,* that will be recognized or produced by the circuit.

The *resolution* of a D/A or A/D IC is usually specified in terms of the bits in the digital code or in terms of the least significant bit (LSB) of the system. An n-bit code allows for 2^n quantization levels, or $2^n - 1$ steps between quantization levels. As the number of bits increases, the step size between quantization levels decreases, therefore increasing the accuracy of the system when a conversion is made between an analog and digital signal. The system resolution can be specified also as the voltage step size between quantization levels. For A/D circuits, the resolution is the smallest input voltage that is detected by the system.

The *speed* of a D/A or A/D converter is determined by the time it takes to perform the conversion process. For D/A converters, the speed is specified as the *settling time.* For A/D converters, the speed is specified as the *conversion time.* The settling time for D/A converters will vary with supply voltage and transition in the digital code; thus, it is specified in the data sheet with the appropriate conditions stated.

A/D converters have a maximum *sampling rate* that limits the speed at which they can perform continuous conversions. The sampling rate is the number of times per second that the analog signal can be sampled and converted into a digital code. For proper A/D conversion, the minimum sampling rate must be at least two times the highest frequency of the analog signal being sampled to satisfy the Nyquist sampling criterion. The conversion speed and other timing factors must be taken into consideration to determine the maximum sampling rate of an A/D converter. **Nyquist A/D converters** use a sampling rate that is slightly more than twice the highest frequency in the analog signal. **Oversampling A/D converters** use sampling rates of N times rate, where N typically ranges from 2 to 64.

Both D/A and A/D converters require a voltage reference in order to achieve absolute conversion accuracy. Some conversion ICs have internal voltage references, whereas others accept external voltage references. For high-performance systems, an external precision reference is needed to ensure long-term stability, load regulation, and control over temperature fluctuations. External precision voltage reference ICs can be found in manufacturer's data books.

Measurement accuracy is specified by the converter's *linearity. Integral linearity* is a measure of linearity over the entire conversion range. It is often defined as the deviation from a straight line drawn between the endpoints and through zero (or the offset value) of the conversion range. Integral linearity is also referred to as *relative accuracy.* The *offset* value is the reference level required to establish the zero or midpoint of the conversion range. *Differential linearity* is the linearity between code transitions. Differential linearity is a measure of the *monotonicity* of the converter. A converter is said to be monotonic if increasing input values result in increasing output values.

Samsung Exhibit 1022, Page 25 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The accuracy and linearity values of a converter are specified in the data sheet in units of the LSB of the code. The linearity can vary with temperature, and so the values are often specified at +25°C as well as over the entire temperature range of the device.

D/A Conversion Processes

Digital codes are typically converted to analog voltages by assigning a voltage weight to each bit in the digital code and then summing the voltage weights of the entire code. A general D/A converter consists of a network of precision resistors, input switches, and level shifters to activate the switches to convert a digital code to an analog current or voltage. D/A ICs that produce an analog current output usually have a faster settling time and better linearity than those that produce a voltage output. When the output current is available, the designer can convert this to a voltage through the selection of an appropriate output amplifier to achieve the necessary response speed for the given application.

D/A converters commonly have a fixed or variable reference level. The reference level determines the switching threshold of the precision switches that form a controlled impedance network, which in turn controls the value of the output signal. **Fixed reference D/A converters** produce an output signal that is proportional to the digital input. **Multiplying D/A** converters produce an output signal that is proportional to the product of a varying reference level times a digital code.

D/A converters can produce bipolar, positive, or negative polarity signals. A four-quadrant multiplying D/A converter allows both the reference signal and the value of the binary code to have a positive or negative polarity. The four-quadrant multiplying D/A converter produces bipolar output signals.

D/A Converter ICs

Most D/A converters are designed for general-purpose control applications. Some D/A converters, however, are designed for special applications, such as video or graphic outputs, high-definition video displays, ultra high-speed signal processing, digital video tape recording, digital attenuators, or high-speed function generators.

D/A converter ICs often include special features that enable them to be interfaced easily to microprocessors or other systems. Microprocessor control inputs, input latches, buffers, input registers, and compatibility to standard logic families are features that are readily available in D/A ICs. In addition, the ICs usually have laser-trimmed precision resistors to eliminate the need for user trimming to achieve full-scale performance.

A/D Conversion Processes

Analog signals can be converted to digital codes by many methods, including integration, **successive approximation**, parallel (flash) conversion, **delta modulation, pulse code modulation**, and **sigma–delta conversion**. Two of the most common A/D conversion processes are successive approximation A/D conversion and parallel or **flash A/D** conversion. Very high-resolution digital audio or video systems require specialized A/D techniques that often incorporate one of these general techniques as well as specialized A/D conversion processes. Examples of specialized A/D conversion techniques are pulse code modulation (PCM), and sigma–delta conversion. PCM is a common voice encoding scheme used not only by the audio industry in digital audio recordings but also by the telecommunications industry for voice encoding and multiplexing. Sigma–delta

Samsung Exhibit 1022, Page 26 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

conversion is an oversampling A/D conversion where signals are sampled at very high frequencies. It has very high resolution and low distortion and is being used in the digital audio recording industry.

Successive approximation A/D conversion is a technique that is commonly used in medium- to high-speed data acquisition applications. It is one of the fastest A/D conversion techniques that requires a minimum amount of circuitry. The conversion times for successive approximation A/D conversion typically range from 10 to 300 μs for 8-b systems.

The successive approximation A/D converter can approximate the analog signal to form an n-bit digital code in n steps. The successive approximation register (SAR) individually compares an analog input voltage to the midpoint of one of n ranges to determine the value of 1 b. This process is repeated a total of n times, using n ranges, to determine the n bits in the code. The comparison is accomplished as follows. The SAR determines if the analog input is above or below the midpoint and sets the bit of the digital code accordingly. The SAR assigns the bits beginning with the most significant bit. The bit is set to a 1 if the analog input is greater than the midpoint voltage, or it is set to a 0 if it is less than the midpoint voltage. The SAR then moves to the next bit and sets it to a 1 or a 0 based on the results of comparing the analog input with the midpoint of the next allowed range. Because the SAR must perform one approximation for each bit in the digital code, an n-bit code requires n approximations.

A successive approximation A/D converter consists of four functional blocks, as shown in Fig. 9.1: the SAR, the analog comparator, a D/A converter, and a clock.

Parallel or flash A/D conversion is used in high-speed applications such as video signal processing, medical imaging, and radar detection systems. A flash A/D converter simultaneously compares the input analog voltage to $2^n - 1$ threshold voltages to produce an n-bit digital code representing the analog voltage. Typical flash A/D converters with 8-b resolution operate at 20–100 MHz.

The functional blocks of a flash A/D converter are shown in Fig. 9.2. The circuitry consists of a precision resistor ladder network, $2^n - 1$ analog comparators, and a digital priority encoder. The resistor network establishes threshold voltages for each allowed quantization level. The analog comparators indicate whether or not the input analog voltage is above or below the threshold at each level. The output of the analog comparators is input to the digital priority encoder. The priority encoder produces the final digital output code that is stored in an output latch.

An 8-b flash A/D converter requires 255 comparators. The cost of high-resolution A/D comparators escalates as the circuit complexity increases and as the number of analog converters rises by $2^n - 1$. As a low-cost alternative, some manufacturers produce modified flash A/D converters that perform the A/D conversion in two steps to reduce the amount of circuitry required. These modified flash A/D converters are also referred to as *half-flash* A/D converters, since they perform only half of the conversion simultaneously.

A/D Converter ICs

A/D converter ICs can be classified as general-purpose, high-speed, flash, and sampling A/D converters. The *general-purpose A/D converters* are typically low speed and low cost, with conversion times ranging from 2 μs to 33 ms. A/D conversion techniques used by these devices typically include successive approximation, tracking, and integrating. The general-purpose A/D converters often have control signals for simplified microprocessor interfacing. These ICs are appropriate for many process control, industrial, and instrumentation applications, as well as for environmental monitoring such as seismology, oceanography, meteorology, and pollution monitoring.

Samsung Exhibit 1022, Page 27 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

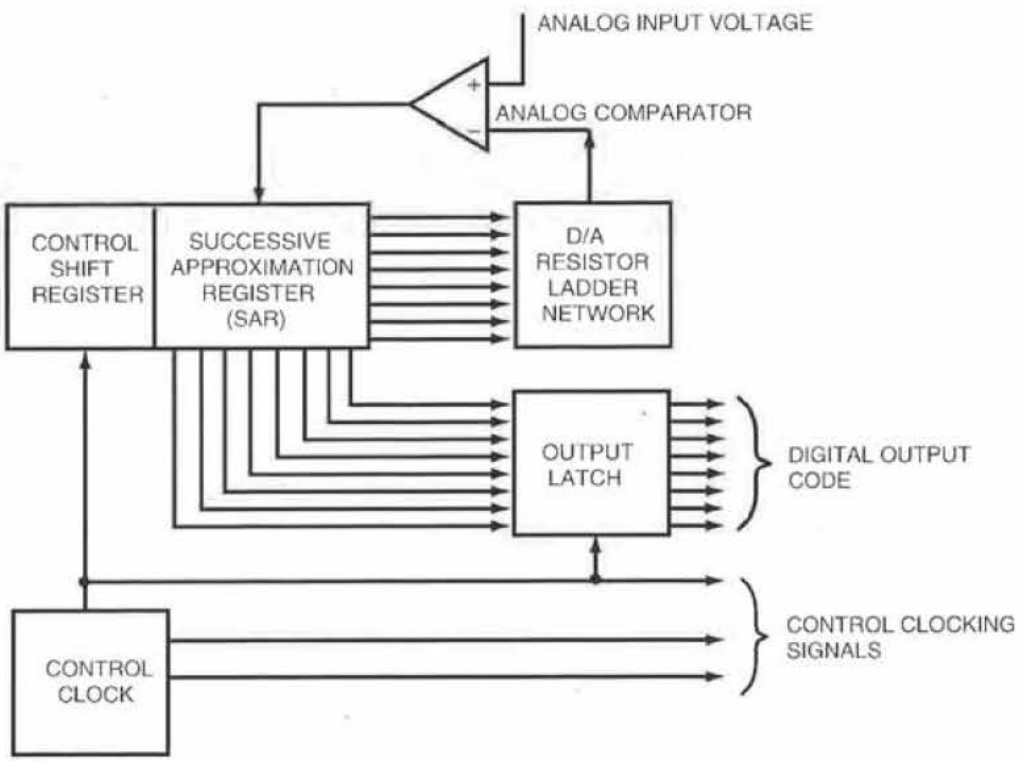


FIGURE 9.1 Successive approximation A/D converter block diagram. (*Source:* Garrod, S. and Borns, R. 1991. *Digital Logic: Analysis, Application, and Design*, p. 919. Copyright ©1991 by Saunders College Publishing, Philadelphia, PA. Reprinted by permission of the publisher.)

High-speed A/D converters have conversion times typically ranging from 400 ns to 3 μs. The higher speed performance of these devices is achieved by using the successive approximation technique, modified flash techniques, and statistically derived A/D conversion techniques. Applications appropriate for these A/D ICs include fast Fourier transform (FFT) analysis, radar digitization, medical instrumentation, and multiplexed data acquisition. Some ICs have been manufactured with an extremely high degree of linearity, to be appropriate for specialized applications in digital spectrum analysis, vibration analysis, geological research, sonar digitizing, and medical imaging.

Flash A/D converters have conversion times ranging typically from 10 to 50 ns. Flash A/D conversion techniques enable these ICs to be used in many specialized high-speed data acquisition applications such as TV video digitizing (encoding), radar analysis, transient analysis, high-speed digital oscilloscopes, medical ultrasound imaging, high-energy physics, and robotic vision applications.

Sampling A/D converters have a sample-and-hold amplifier circuit built into the IC. This eliminates the need for an external sample-and-hold circuit. The throughput of these A/D converter ICs ranges typically from 35 kHz to 100 MHz. The speed of the system is dependent on the A/D technique used by the sampling A/D converter.

A/D converter ICs produce digital codes in a serial or parallel format, and some ICs offer the designer both formats. The digital outputs are compatible with standard logic families to facilitate

Samsung Exhibit 1022, Page 28 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

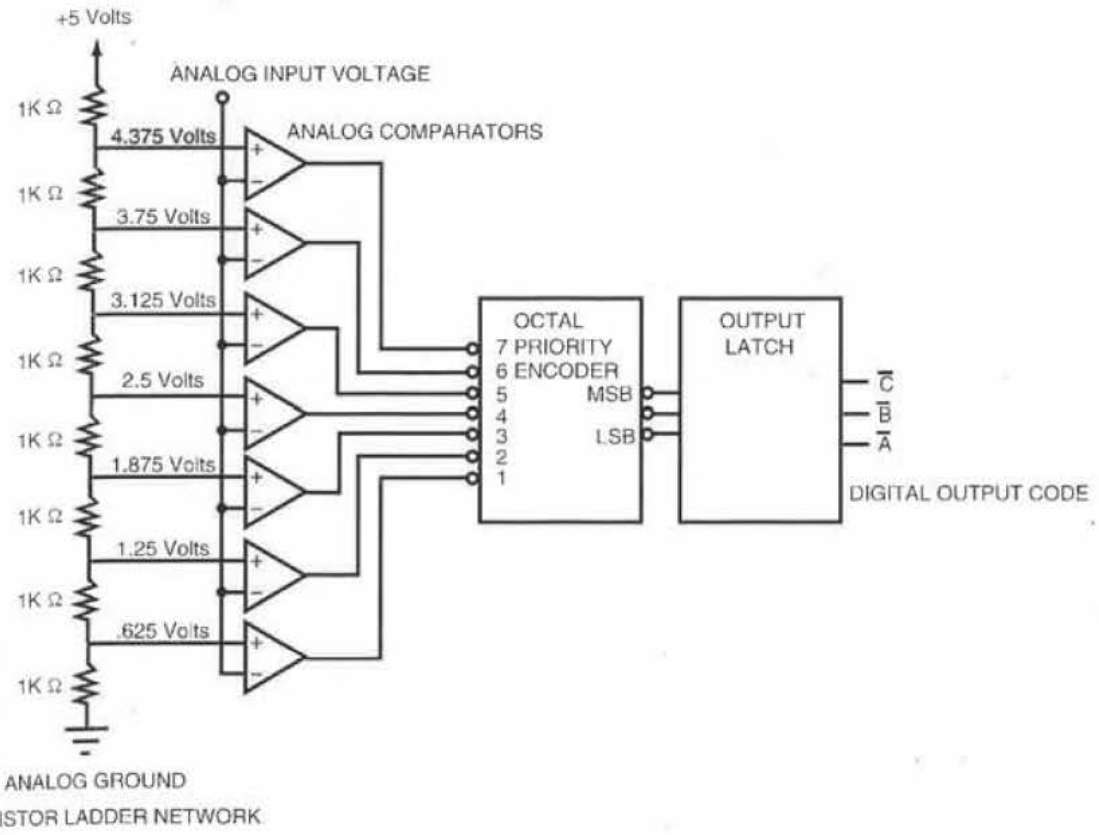


FIGURE 9.2 Flash A/D converter block diagram. (*Source:* Garrod, S. and Borns, R. *Digital Logic: Analysis, Application, and Design*, p. 928. Copyright ©1991 by Saunders College Publishing, Philadelphia, PA. Reprinted by permission of the publisher.)

interfacing to other digital systems. In addition, some A/D converter ICs have a built-in analog multiplexer and therefore can accept more than one analog input signal.

Pulse code modulation (PCM) ICs are high-precision A/D converters. The PCM IC is often refered to as a PCM *codec* with both encoder and decoder functions. The encoder portion of the codec performs the A/D conversion, and the decoder portion of the codec performs the D/A conversion. The digital code is usually formatted as a serial data stream for ease of interfacing to digital transmission and multiplexing systems.

PCM is a technique where an analog signal is sampled, quantized, and then encoded as a digital word. The PCM IC can include successive approximation techniques or other techniques to accomplish the PCM encoding. In addition, the PCM codec may employ nonlinear data compression techniques, such as companding, if it is necessary to minimize the number of bits in the output digital code. Companding is a logarithmic technique used to compress a code to fewer bits before transmission. The inverse logarithmic function is then used to expand the code to its original number of bits before converting it to the analog signal. Companding is typically used in telecommunications transmission systems to minimize data transmission rates without degrading the resolution of low-amplitude signals. Two standardized companding techniques are used extensively: A-law and μ-law. The A-law companding is used in Europe, whereas the μ-law is used predominantly in the United States and Japan. Linear PCM conversion is used in

Samsung Exhibit 1022, Page 29 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

high-fidelity audio systems to preserve the integrity of the audio signal throughout the entire analog range.

Digital signal processing (DSP) techniques provide another type of A/D conversion ICs. Specialized A/D conversion such as *adaptive differential pulse code modulation* (ADPCM), sigma–delta modulation, *speech subband encoding, adaptive predictive speech encoding,* and *speech recognition* can be accomplished through the use of DSP systems. Some DSP systems require analog front ends that employ traditional PCM codec ICs or DSP interface ICs. These ICs can interface to a digital signal processor for advanced A/D applications. Some manufacturers have incorporated DSP techniques on board the single-chip A/D IC, as in the case of the DSP56ACD16 sigma–delta modulation IC by Motorola.

Integrating A/D converters are used for conversions that must take place over a long period of time, such as digital voltmeter applications or sensor applications such as thermocouples. The integrating A/D converter produces a digital code that represents the average of the signal over time. Noise is reduced by means of the signal averaging, or integration. Dual-slope integration is accomplished by a counter that advances while an input voltage charges a capacitor in a specified time interval, T. This is compared to another count sequence that advances while a reference voltage is discharging across the same capacitor in a time interval, *delta t*. The ratio of the charging count value to the discharging count value is proportional to the ratio of the input voltage to the reference voltage. Hence, the integrating converter provides a digital code that is a measure of the input voltage averaged over time. The conversion accuracy is independent of the capacitor and the clock frequency since they affect both the charging and discharging operations. The charging period, T, is selected to be the period of the fundamental frequency to be rejected. The maximum conversion rate is slightly less than $1/(2\,T)$ conversions per second. While this limits the conversion rate to be too slow for high-speed data acquisition applications, it is appropriate for long-duration applications of slowly varying input signals.

Grounding and Bypassing on D/A and A/D ICs

D/A and A/D converter ICs require correct grounding and capacitive bypassing in order to operate according to performance specifications. The digital signals can severely impair analog signals. To combat the electromagnetic interference induced by the digital signals, the analog and digital grounds should be kept separate and should have only one common point on the circuit board. If possible, this common point should be the connection to the power supply.

Bypass capacitors are required at the power connections to the IC, the reference signal inputs, and the analog inputs to minimize noise that is induced by the digital signals. Each manufacturer specifies the recommended bypass capacitor locations and values in the data sheet. The 1-μF tantalum capacitors are commonly recommended, with additional high-frequency power supply decoupling sometimes being recommended through the use of ceramic disc shunt capacitors. The manufacturers' recommendations should be followed to ensure proper performance.

Selection Criteria for D/A and A/D Converter ICs

Hundreds of D/A and A/D converter ICs are available, with prices ranging from a few dollars to several hundred dollars each. The selection of the appropriate type of converter is based on the application requirements of the system, the performance requirements, and cost. The following issues should be considered in order to select the appropriate converter.

Samsung Exhibit 1022, Page 30 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

1. What are the input and output requirements of the system? Specify all signal current and voltage ranges, logic levels, input and output impedances, digital codes, data rates, and data formats.
2. What level of accuracy is required? Determine the resolution needed throughout the analog voltage range, the dynamic response, the degree of linearity, and the number of bits encoding.
3. What speed is required? Determine the maximum analog input frequency for sampling in an A/D system, the number of bits for encoding each analog signal, and the rate of change of input digital codes in a D/A system.
4. What is the operating environment of the system? Obtain information on the temperature range and power supply to select a converter that is accurate over the operating range.

Final selection of D/A and A/D converter ICs should be made by consulting manufacturers to obtain their technical specifications of the devices. Major manufacturers of D/A and A/D converters include Analog Devices, Burr–Brown, DATEL, Maxim, National, Phillips Components, Precision Monolithics, Signetics, Sony, Texas Instruments, Ultra Analog, and Yamaha. Information on contacting these manufacturers and others can be found in an IC *Master Catalog*.

Defining Terms

Companding: A process designed to minimize the transmission bit rate of a signal by compressing it prior to transmission and expanding it upon reception. It is a rudimentary "data compression" technique that requires minimal processing.

Delta modulation: An A/D conversion process where the digital output code represents the change, or slope, of the analog input signal, rather than the absolute value of the analog input signal. A 1 indicates a rising slope of the input signal. A 0 indicates a falling slope of the input signal. The sampling rate is dependent on the derivative of the signal, since a rapidly changing signal would require a rapid sampling rate for acceptable performance.

Fixed reference D/A converter: The analog output is proportional to a fixed (nonvarying) reference signal.

Flash A/D: The fastest A/D conversion process available to date, also referred to as parallel A/D conversion. The analog signal is simultaneously evaluated by $2^n - 1$ comparators to produce an n-bit digital code in one step. Because of the large number of comparators required, the circuitry for flash A/D converters can be very expensive. This technique is commonly used in digital video systems.

Integrating A/D: The analog input signal is integrated over time to produce a digital signal that represents the area under the curve, or the integral.

Multiplying D/A: A D/A conversion process where the output signal is the product of a digital code multiplied times an analog input reference signal. This allows the analog reference signal to be scaled by a digital code.

Nyquist A/D converters: A/D converters that sample analog signals that have a maximum frequency that is less than the Nyquist frequency. The Nyquist frequency is defined as one-half of the sampling frequency. If a signal has frequencies above the Nyquist frequency, a distortion called *aliasing* occurs. To prevent aliasing, an *antialiasing filter* with a flat passband and very sharp rolloff is required.

Oversampling converters: A/D converters that sample frequencies at a rate much higher than the Nyquist frequency. Typical oversampling rates are 32 and 64 times the sampling rate that would be required with the Nyquist converters.

Samsung Exhibit 1022, Page 31 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Pulse code modulation (PCM): An A/D conversion process requiring three steps: the analog signal is sampled, quantized, and encoded into a fixed length digital code. This technique is used in many digital voice and audio systems. The reverse process reconstructs an analog signal from the PCM code. The operation is very similar to other A/D techniques, but specific PCM circuits are optimized for the particular voice or audio application.

Sigma–delta A/D conversion: An *oversampling* A/D conversion process where the analog signal is sampled at rates much higher (typically 64 times) than the sampling rates that would be required with a Nyquist converter. Sigma–delta modulators integrate the analog signal before performing the delta modulation. The integral of the analog signal is encoded rather than the change in the analog signal, as is the case for traditional delta modulation. A digital sample rate reduction filter (also called a digital decimation filter) is used to provide an output sampling rate at twice the Nyquist frequency of the signal. The overall result of oversampling and digital sample rate reduction is greater resolution and less distortion compared to a Nyquist converter process.

Successive approximation: An A/D conversion process that systematically evaluates the analog signal in n steps to produce an n-bit digital code. The analog signal is successively compared to determine the digital code, beginning with the determination of the most significant bit of the code.

References

Analog Devices. 1989. *Analog Devices Data Conversion Products Data Book.* Analog Devices, Inc., Norwood, MA.

Burr–Brown. 1989. *Burr–Brown Integrated Circuits Data Book.* Burr–Brown, Tucson, AZ.

DATEL. 1988. *DATEL Data Conversion Catalog.* DATEL, Inc., Mansfield, MA.

Drachler, Will, and Bill Murphy. 1995. New High-Speed, Low-Power Data-Acquisition ICs. *Analog Dialogue* 29(2):3–6. Analog Devices, Inc., Norwood, MA.

Garrod, S. and Borns, R. 1991. *Digital Logic: Analysis, Application and Design*, Chap. 16. Saunders College Publishing, Philadelphia, PA.

Jacob, J.M. 1989. *Industrial Control Electronics*, Chap. 6. Prentice–Hall, Englewood Cliffs, NJ.

Keiser, B. and Strange, E. 1995. *Digital Telephony and Network Integration*, 2nd ed. Van Nostrand Reinhold, New York.

Motorola. 1989. *Motorola Telecommunications Data Book.* Motorola, Inc., Phoenix, AZ.

National Semiconductor. 1989. *National Semiconductor Data Acquisition Linear Devices Data Book.* National Semiconductor Corp., Santa Clara, CA.

Park, S. 1990. *Principles of Sigma–Delta Modulation for Analog-to-Digital Converters.* Motorola, Inc., Phoenix, AZ.

Texas Instruments. 1986. *Texas Instruments Digital Signal Processing Applications with the TMS320 Family.* Texas Instruments, Dallas, TX.

Texas Instruments. 1989. *Texas Instruments Linear Circuits Data Acquisition and Conversion Data Book.* Texas Instruments, Dallas, TX.

Further Information

Analog Devices, Inc. has edited or published several technical handbooks to assist design engineers with their data acquisition system requirements. These references should be consulted for extensive technical information and depth. The publications include *Analog-Digital Conversion Handbook*, by the engineering staff of Analog Devices, published by Prentice–Hall, Englewood Cliffs, NJ, 1986; *Nonlinear Circuits Handbook, Transducer Interfacing Hand-*

Samsung Exhibit 1022, Page 32 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

book, and *Synchro and Resolver Conversion*, all published by Analog Devices Inc., Norwood, MA.

Engineering trade journals and design publications often have articles describing recent A/D and D/A circuits and their applications. These publications include *EDN Magazine*, *EE Times*, and *IEEE Spectrum*. Research-related topics are covered in *IEEE Transactions on Circuits and Systems* and also *IEEE Transactions on Instrumentation and Measurement*.

Samsung Exhibit 1022, Page 33 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

10

Signal Space

10.1 Introduction 118
10.2 Fundamentals 118
10.3 Application of Signal Space Representation to Signal Detection 125
10.4 Application of Signal Space Representation to Parameter Estimation 126
Wavelet Transforms • Mean Square Estimation—the Orthogonality Principle

Rodger E. Ziemer
University of Colorado at Colorado Springs

10.1 Introduction

Signal space concepts have their roots in the mathematical theory of inner product spaces known as **Hilbert spaces** [Stakgold, 1967]. Many books on linear systems touch on the subject of signal spaces in the context of Fourier series and transforms [Ziemer, Tranter, and Fannin, 1993; Frederick and Carlson, 1971]. The applications of signal space concepts in communication theory find their power in the representation of signal detection and estimation problems in geometrical terms, which provides much insight into signalling techniques and communication system design. The first person to have apparently exploited the power of signal space concepts in communication theory was the Russian Kotel'nikov [1968] who presented his doctoral dissertation in January, 1947. Wozencraft and Jacobs [1965] expanded on this approach and is still today widely referenced. Arthurs and Dym [1962] made use of signal space concepts in the performance analysis of several digital modulation schemes. A one-chapter summary of the use of signal space methods in signal detection and estimation is provided in Ziemer and Tranter [1995]. Another application of signal space concepts is in signal and image compression. Wavelet theory [Rioul and Vetterli, 1991] is currently finding use in these application areas. In the next section, the fundamentals of generalized vector spaces are summarized, followed by an overview of several applications to signal representations.

10.2 Fundamentals

A linear space or **vector space** (signal space) [Stakgold, 1967] is a collection of elements (called **vectors**) $x, y, z, \ldots$, for which the following axioms are satisfied:

1. To every pair of vectors x and y there corresponds a vector $x + y$, with the properties:
   a. $x + y = y + x$
   b. $x + (y + z) = (x + y) + z$

0-8493-8349-8/97/$0.00+$.50
© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 34 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

c. There exists a unique element 0 such that $\boldsymbol{x} + 0 = \boldsymbol{x}$ for every $\boldsymbol{x}$
d. To every $\boldsymbol{x}$, there exists a unique vector labeled $-\boldsymbol{x}$ such that $\boldsymbol{x} + (-\boldsymbol{x}) = 0$

2. To all vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, and all numbers α and β (in general, complex), the following commutative and associative rules hold:

a. $\alpha(\beta\boldsymbol{x}) = (\alpha\beta)\boldsymbol{x}$
b. $(\alpha + \beta)\boldsymbol{x} = \alpha\boldsymbol{x} + \beta\boldsymbol{y}$
c. $\alpha(\boldsymbol{x} + \boldsymbol{y}) = \alpha\boldsymbol{x} + \alpha\boldsymbol{y}$
d. $1\boldsymbol{x} = \boldsymbol{x}$, where 1 is the **identity element**

A vector is said to be a **linear combination** of the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ in a vector space if there exist numbers (in general, complex) $\alpha_1, \alpha_2, \ldots, \alpha_k$ such that

$$\boldsymbol{x} = \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i \tag{10.1}$$

The vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ are said to be **linearly dependent** (or form a *dependent set*) if there exist complex numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$, not all zero, such that

$$\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_k\boldsymbol{x}_k = 0 \tag{10.2}$$

If Eq. (10.2) can be satisfied only for $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, the vectors are **linearly independent**.

One is tempted to use the infinite-sum version of Eq. (10.2) in defining the notion of independence for an infinite set of vectors. This is not true in general; one needs the notion of convergence, which is based on the concept of distance between vectors.

With the idea of linear independence firmly in mind, the concept of **dimension** of a vector space readily follows. A vector space is *n dimensional* if it possesses a set of n **independent vectors**, but every set of $n + 1$ vectors is a dependent set. If for every positive integer k, a set of k independent vectors in the space can be found, the space is said to be infinite dimensional. By a **basis** for a vector space is meant a finite set of vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ with the following attributes:

1. They are linearly independent.
2. Every vector $\boldsymbol{x}$ in the space can be written as a linear combination of the basis vectors; that is

$$\boldsymbol{x} = \sum_{i=1}^{k} \xi_i \boldsymbol{e}_i \tag{10.3}$$

It can be proved that the representation (10.3) is unique and that if the space is n dimensional, any set of n independent vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ forms a basis.

The next concept to be developed is that of a **metric space**. In addition to the addition of vectors and multiplication of vectors by scalars, as is true of ordinary three-dimensional vectors, it is important to have the notions of length and direction of a vector imposed. In other words, a metric structure must be added to the algebraic structure already defined. A collection of elements $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots$ in a space will be called a metric space if to each pair of elements $\boldsymbol{x}, \boldsymbol{y}$ there corresponds a real number $d(\boldsymbol{x}, \boldsymbol{y})$ satisfying the properties:

1. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$
2. $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ with equality if and only if $\boldsymbol{x} = \boldsymbol{y}$
3. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{x}, \boldsymbol{z})$ (called the **triangle inequality**)

The function $d(\boldsymbol{x}, \boldsymbol{y})$ is called a **metric** (or distance function). Note that the elements in the space *need not be vectors*; there may not be any way of adding elements or multiplying them by scalars as required for a vector space.

Samsung Exhibit 1022, Page 35 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

With the definition of a metric, one can now discuss the idea of convergence of a sequence $\{\boldsymbol{x}_k\}$ of elements in the space. Note that $d(\boldsymbol{x}_k, \boldsymbol{x})$ is a sequence of real numbers. Therefore, it is sensible to write that

$$\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x} \tag{10.4}$$

if the sequence of numbers $d(\boldsymbol{x}_k, \boldsymbol{x})$ converges to 0 in the ordinary sense of convergence of sequences of real numbers. If

$$\lim_{m,p\to\infty} d(\boldsymbol{x}_m, \boldsymbol{x}_p) = 0 \tag{10.5}$$

the sequence $\{\boldsymbol{x}_k\}$ is said to be a **Cauchy sequence**. It can be shown that if a sequence $\{\boldsymbol{x}_k\}$ converges, it is a Cauchy sequence. The converse is not necessarily true, for the limit may have carelessly been excluded from the space. If the converse is true, then the metric space is said to be **complete**.

The next vector space concept to be defined is that of length or norm of a vector. A **normed vector space** (or linear space) is a vector space in which a real-valued function $\|\boldsymbol{x}\|$ (known as the *norm* of $\boldsymbol{x}$) is defined, with the properties

1. $\|\boldsymbol{x}\| \geq 0$ with equality if and only if $\boldsymbol{x} = 0$
2. $\|\alpha\boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|$
3. $\|\boldsymbol{x}_1 + \boldsymbol{x}_2\| \leq \|\boldsymbol{x}_1\| + \|\boldsymbol{x}_2\|$

A normed vector space is automatically a metric space if the metric is defined as

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\| = \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y}\rangle^{1/2} \tag{10.6}$$

(see definition below) which is called the natural metric for the space. A normed vector space may be viewed either as a linear space, a metric space, or both. Its elements may be interpreted as *vectors* or *points*.

The structure of a normed vector space will now be refined further with the definition of the notion of angle between two vectors. In particular, it will be possible to tell whether two vectors are perpendicular. The notion of angle between two vectors will be obtained by defining the **inner product** (also known as a scalar or dot product). In general, an inner product in a vector space is a complex-valued function of ordered pairs $\boldsymbol{x}$, $\boldsymbol{y}$ with the properties

1. $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \langle \boldsymbol{y}, \boldsymbol{x}\rangle^*$ (the asterisk denotes complex conjugate)
2. $\langle \alpha\boldsymbol{x}, \boldsymbol{y}\rangle = \alpha\langle \boldsymbol{x}, \boldsymbol{y}\rangle$
3. $\langle \boldsymbol{x}_1 + \boldsymbol{x}_2, \boldsymbol{y}\rangle = \langle \boldsymbol{x}_1, \boldsymbol{y}\rangle + \langle \boldsymbol{x}_2, \boldsymbol{y}\rangle$
4. $\langle \boldsymbol{x}, \boldsymbol{x}\rangle \geq 0$ with equality if and only if $\boldsymbol{x} = 0$

From the first two properties, it follows that

$$\langle \boldsymbol{x}, \alpha\boldsymbol{y}\rangle = \alpha^*\langle \boldsymbol{x}, \boldsymbol{y}\rangle \tag{10.7}$$

Also, **Schwarz's inequality** can be proved and is given by

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \langle \boldsymbol{x}, \boldsymbol{x}\rangle^{1/2}\langle \boldsymbol{y}, \boldsymbol{y}\rangle^{1/2} \tag{10.8}$$

with equality if and only if $\boldsymbol{x} = \alpha\boldsymbol{y}$. The real, nonnegative quantity $\langle \boldsymbol{x}, \boldsymbol{x}\rangle^{1/2}$ satisfies all of the properties of a norm. Therefore, it is adopted as the definition of the norm, and Schwartz's inequality assumes the form

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|\,\|\boldsymbol{y}\| \tag{10.9}$$

Samsung Exhibit 1022, Page 36 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The natural metric in the space is given by Eq. (10.6). An inner product space, which is complete in its natural metric, is called a Hilbert space.

Example 10.1

Consider the space of all complex-valued functions $x(t)$ defined on $a \leq t \leq b$ for which the integral

$$E_x = \int_a^b |x(t)|^2 \, dt \tag{10.10}$$

exists (i.e., the space of all finite-energy signals in the interval $[a, b]$). The inner product is defined as

$$\langle x, y \rangle = \int_a^b x(t) y^*(t) \, dt \tag{10.11}$$

The natural norm is

$$\|x\| = \left[\int_a^b |x(t)|^2 \, dt \right]^{1/2} \tag{10.12}$$

and the metric is

$$d(x, y) = \|x - y\| = \left[\int_a^b |x(t) - y(t)|^2 \, dt \right]^{1/2} \tag{10.13}$$

respectively. Schwarz's inequality becomes

$$\left| \int_a^b |x(t) y^*(t)| \, dt \right| \leq \left[\int_a^b |x(t)|^2 \, dt \right]^{1/2} \left[\int_a^b |y(t)|^2 \, dt \right]^{1/2} \tag{10.14}$$

It can be shown that this space is complete and, hence, is a Hilbert space.

An additional requirement that can be imposed on a Hilbert space is separability, which, roughly speaking, restricts the number of elements in the space. A Hilbert space $\mathcal{H}$ is **separable** if there exists a **countable** (i.e., can be put in one-to-one correspondence with the positive integers) set of elements $(f_1, f_2, \ldots, f_n, \ldots)$ whose finite linear combinations are such that for any element f in $\mathcal{H}$ there exist an index N and constants $\alpha_1, \alpha_2, \ldots, \alpha_N$ such that

$$\left\| f - \sum_{k=1}^{N} \alpha_k f_k \right\| < \epsilon \tag{10.15}$$

The set $(f_1, f_2, \ldots, f_n, \ldots)$ is called a **spanning set**. The discussions from here on are limited to separable Hilbert spaces.

Any finite-dimensional Hilbert space E_n is separable. In fact, there exists a set of n vectors $(f_1, f_2, \ldots, f_n)$ such that each vector x in E_n has the representation

$$x = \sum_{k=1}^{n} \alpha_k f_k \tag{10.16}$$

It can be shown that the spaces consisting of square-integrable functions on the intervals $[a, b]$, $(-\infty, b]$, $[a, \infty)$, and $(-\infty, \infty)$ are all separable, where a and b are finite, and spanning sets exist

Samsung Exhibit 1022, Page 37 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

for each of these spaces. For example, a spanning set for the space of square-integrable functions on $[a, b]$ is the set $(1, t, t^2, \ldots)$, which is clearly countable.

The concepts of convergence and Cauchy sequences carry over to Hilbert spaces, with *convergence in the mean* defined as

$$\lim_{k\to\infty} \int_a^b |\boldsymbol{x}_k(t) - \boldsymbol{x}(t)|^2 = 0 \tag{10.17}$$

Similarly, the ideas of independence and basis sets apply to Hilbert spaces in infinite-dimensional form.

It is necessary to distinguish between the concepts of a *basis* and a *spanning* set consisting of independent vectors. As ε is reduced in Eq. (10.15), it is expected that N must be increased, and it may also be necessary to change the previously found coefficients $\alpha_1, \ldots, \alpha_N$. Hence, there might not exist a fixed sequence of constants $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ with the property

$$\boldsymbol{x} = \sum_{k=1}^{\infty} \xi_k \boldsymbol{f}_k \tag{10.18}$$

as would be required if the set $\{\boldsymbol{f}_k\}$ were a basis. For example, on the space of square-integrable functions on $[-1, 1]$, the independent set $\boldsymbol{f}_0 = 1, \boldsymbol{f}_1 = t, \boldsymbol{f}_2 = t^2, \ldots$, is a spanning set, but not a basis, since there are many square-integrable functions on $[-1, 1]$ that cannot be expanded in a series, like Eq. (10.19) (an example is $|t|$). Odd as it may seem at first, it is possible if the powers of t in this spanning set are regrouped into the set of polynomials known as the Legendre polynomials, $P_k(t)$.

Two vectors $\boldsymbol{x}, \boldsymbol{y}$ are **orthogonal** or perpendicular if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$. A finite or countably infinite set of vectors $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_k, \ldots\}$ is said to be an *orthogonal set* if $\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle = 0, i \neq j$. A *proper orthogonal set* is an orthogonal set none of whose elements is the zero vector. A proper orthogonal set is an independent set. A set is **orthonormal** if

$$\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \tag{10.19}$$

An important concept is that of a **linear manifold** in a Hilbert space. A set M is said to be a linear manifold if, for $\boldsymbol{x}$ and $\boldsymbol{y}$ belonging to M, so does $\alpha\boldsymbol{x} + \beta\boldsymbol{y}$ for arbitrary complex numbers α and β; thus, M is itself a linear space. If a linear manifold is a closed set, it is called a closed linear manifold (i.e., every Cauchy sequence has a limit in the space) and is itself a Hilbert space. In three-dimensional Euclidean space, linear manifolds are simply lines and planes containing the origin. In a finite-dimensional space, every linear manifold is necessarily closed.

Let M be a linear manifold, closed or not. Consider the set $M^\perp$ of all vectors which are orthogonal to every vector in M. It is a linear manifold, which can be shown to be closed. If M is closed, M and $M^\perp$ are known as **orthogonal complements**. Given a linear manifold M, each vector in the space can be decomposed in a unique manner, as a sum $\boldsymbol{x}_p + \boldsymbol{z}$, where $\boldsymbol{x}_p$ is in M and $\boldsymbol{z}$ is in $M^\perp$.

Given an infinite orthonormal set $\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n, \ldots\}$ in the space of all square-integrable functions on the interval $[a, b]$, let $\{a_n\}$ be a sequence of complex numbers. The **Riesz–Fischer theorem** tells how to represent an element in the space and states:

1. If

$$\sum_{n=1}^{\infty} |a_n|^2 \tag{10.20}$$

Samsung Exhibit 1022, Page 38 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

diverges, then

$$\sum_{n=1}^{\infty} a_n \phi_n \tag{10.21}$$

diverges.

2. If Eq. (10.20) converges, then Eq. (10.21) also converges to some element $\boldsymbol{g}$ in the space and

$$a_n = \langle \boldsymbol{g}, \phi_n \rangle \tag{10.22}$$

The next question that arises is how to construct an orthonormal set from an independent set $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k, \ldots\}$. A way to do this is known as the **Gram–Schmidt procedure**. The construction is as follows:

1. Pick a function from the set $\{\boldsymbol{e}_k\}$, say $\boldsymbol{e}_1$. Let

$$\phi_1 = \frac{\boldsymbol{e}_1}{\langle \boldsymbol{e}_1, \boldsymbol{e}_1 \rangle^{\frac{1}{2}}} \tag{10.23}$$

2. Remove from a second function in the set $\{\boldsymbol{e}_k\}$, say $\boldsymbol{e}_2$, its projection on ϕ_1. This yields

$$\boldsymbol{g}_2 = \boldsymbol{e}_2 - \langle \boldsymbol{e}_2, \phi_1 \rangle \phi_1 \tag{10.24}$$

The vector $\boldsymbol{g}_2$ is a linear combination of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$, and is orthogonal to ϕ_2. To normalize it, form

$$\phi_2 = \frac{\boldsymbol{g}_2}{\langle \boldsymbol{g}_2, \boldsymbol{g}_2 \rangle^{\frac{1}{2}}} \tag{10.25}$$

3. Pick another function from the set $\{\boldsymbol{e}_k\}$, say $\boldsymbol{e}_3$, and form

$$\boldsymbol{g}_3 = \boldsymbol{e}_3 - \langle \boldsymbol{e}_3, \phi_2 \rangle \phi_2 - \langle \boldsymbol{e}_3, \phi_1 \rangle \phi_1 \tag{10.26}$$

Normalize $\boldsymbol{g}_3$ in a manner similar to that used for $\boldsymbol{g}_2$.

4. Continue until all functions in the set $\{\boldsymbol{e}_k\}$ have been used.

Note that the sets $\{\boldsymbol{e}_k\}$, $\{\boldsymbol{g}_k\}$, and $\{\phi_k\}$ all generate the same linear manifold.

A basis consisting of orthonormal vectors is known as an **orthonormal basis**. If the basis vectors are not normalized, it is simply an *orthogonal basis*.

Example 10.2

Consider the interval $[-1, 1]$ and the independent set $\boldsymbol{e}_0 = 1, \boldsymbol{e}_2 = t, \ldots, \boldsymbol{e}_k = t^k, \ldots$. The Gram–Schmidt procedure applied to this set without normalization, but with the requirement that all orthogonal functions take on the value 1 at $t = 1$, gives the set of *Lengendre polynomials*, which is

$$\psi_0(t) = 1, \qquad \psi_1(t) = t, \qquad \psi_2(t) = \tfrac{1}{2}(3t^2 - 1), \qquad \psi_3(t) = \tfrac{1}{2}(5t^3 - 3t), \ldots \tag{10.27}$$

It is next desired to approximate an arbitrary vector $\boldsymbol{x}$ in a Hilbert space in terms of a linear combination of the independent set $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$, where $k \leq n$ if the space is an n-dimensional Euclidean space, and k is an arbitrary integer if the space is infinite dimensional. First, the orthonormal

Samsung Exhibit 1022, Page 39 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

set $\{\phi_1, \ldots, \phi_k\}$ is constructed from $\{e_1, \ldots, e_k\}$. The unique, best approximation to x is the **Fourier sum**

$$\sum_{i=1}^{k} \langle x, \phi_i \rangle \phi_i \tag{10.28}$$

which is geometrically the projection of x onto the linear manifold generated by $\{\phi_1, \ldots, \phi_k\}$, or equivalently, the sum of the projections along the individual axes defined by $\phi_1, \phi_2, \ldots, \phi_k$. The square of the distance between x and its projection is

$$\left\| x - \sum_{i=1}^{k} \langle x, \phi_i \rangle \phi_i \right\|^2 = \|x\|^2 - \sum_{i=1}^{k} |\langle x, \phi_i \rangle|^2 \tag{10.29}$$

Since the left-hand side is nonnegative, Eq. (10.29) gives **Bessel's inequality**, which is

$$\|x\|^2 \geq \sum_{i=1}^{k} |\langle x, \phi_i \rangle|^2 \tag{10.30}$$

A convenient feature of the Fourier sum is the following: If another vector ϕ_{k+1} is added the orthonormal approximating set of vectors, the best approximation now becomes

$$\sum_{i=1}^{k+1} \langle x, \phi_i \rangle \phi_i \tag{10.31}$$

Thus, an additional term is added to the series expansion without changing previously computed coefficients, which makes the extension to a countably infinite orthonormal approximating set simple to envision. In the case of an infinite orthonormal approximating set, Bessel's inequality (10.30) now has an infinite limit on the sum. Does the approximating sum (10.31) converge to x? The answer is that convergence can be guaranteed only if the set $\{\phi_k\}$ is extensive enough, that is, if it is a *basis or a complete orthonormal set.* In such cases, Eq. (10.30) becomes an equality. In fact, a number of equivalent criteria can be stated to determine whether an orthonormal set $\{\phi_k\}$ is a basis or not [Stakgold, 1967]. These are:

1. In finite n-dimensional Euclidean space, $\{\phi_k\}$ has exactly n elements for completeness.
2. For every x in the space of square-integrable functions

$$x = \sum_{i} \langle x, \phi_i \rangle \phi_i \tag{10.32}$$

3. For every x in the space of square-integrable functions

$$\|x\|^2 = \sum_{i=1}^{k} |\langle x, \phi_i \rangle|^2 \tag{10.33}$$

(known as **Parseval's equality**).

4. The only x in the space of square-integrable functions for which all of the Fourier coefficients vanish is the 0 function.
5. There exists no function $\phi(t)$ in the space of square-integrable functions such that $\{\phi, \phi_1, \phi_2, \ldots, \phi_k, \ldots\}$ is an orthonormal set.

Samsung Exhibit 1022, Page 40 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Examples of complete orthonormal sets of trigonometric functions over the interval $[0, T]$ are as follows: (1) The complex exponentials with frequencies equal to the harmonics of the fundamental frequency $\omega_0 = 2\pi/T$, or

$$\frac{1}{\sqrt{2\pi}}, \quad \frac{e^{j\omega_0 t}}{\sqrt{2\pi}}, \quad \frac{e^{-j\omega_0 t}}{\sqrt{2\pi}}, \quad \frac{e^{j2\omega_0 t}}{\sqrt{2\pi}}, \quad \frac{e^{-j2\omega_0 t}}{\sqrt{2\pi}}, \ldots \tag{10.34}$$

The factor $(2\pi)^{1/2}$ in the denominator is necessary to normalize the functions and is often not included in the definition of the complex exponential Fourier series. (2) The sines and cosines with frequencies equal to harmonics of the fundamental frequency $\omega_0 = 2\pi/T$, or

$$\frac{1}{\sqrt{2\pi}}, \quad \frac{(\cos \omega_0 t)}{\sqrt{\pi}}, \quad \frac{(\sin \omega_0 t)}{\sqrt{\pi}}, \quad \frac{(\cos 2\omega_0 t)}{\sqrt{2\pi}}, \quad \frac{(\sin 2\omega_0 t)}{\sqrt{2\pi}}, \ldots \tag{10.35}$$

Note that if any function is left out of these sets, the basis is incomplete.

10.3 Application of Signal Space Representation to Signal Detection

The M-ary signal detection problem is as follows: given M signals, $s_0(t), s_1(t), \ldots, s_M(t)$, defined over $0 \leq t \leq T$. One is chosen at random and sent each T-second interval through a channel that adds white, Gaussian noise of power spectral density $N_0/2$ to it. The challenge is to design a receiver that will decide which signal was sent through the channel during each T-second interval with minimum probability of making an error.

An approach to this problem, as expanded upon in greater detail by Wozencraft and Jacobs [1965, Chap. 4] and Ziemer and Tranter [1995, Chap. 9], is to construct a linear manifold, called the signal space, using the Gram–Schmidt procedure on the M signals. Suppose that this results in the orthonormal basis set $\{\phi_1, \phi_2, \ldots, \phi_K\}$ where $K \leq M$. The received signal plus noise is represented in this signal space as vectors with coordinates (note that they depend on the signal transmitted)

$$Z_{ij} = A_{ij} + N_j, \qquad \begin{matrix} j = 1, 2, \ldots, K, \\ i = 1, 2, \ldots, M \end{matrix} \tag{10.36}$$

where

$$A_{ij} = \int_0^T s_i(t)\phi_j(t)\, dt \tag{10.37}$$

The numbers Z_{ij} are components of vectors referred to as the **signal vectors**, and the space of all signal vectors is called the **observation space**. An apparent problem with this approach is that not all possible noise waveforms added to the signal can be represented as vectors in this K-dimensional observation space. The part of the noise that is represented is

$$n_1(t) = \sum_{j=1}^{K} N_j \phi_j(t) \tag{10.38}$$

where

$$N_j = \int_0^T n(t)\phi_j(t)\, dt \tag{10.39}$$

In terms of Hilbert space terminology, Eq. (10.38) is the projection of the noise waveform onto the

Samsung Exhibit 1022, Page 41 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Kennedy, R.S. 1969. *Fading Dispersive Communication Channels*, Wiley, New York.

Knowles, M.D. and Drukarev, A.I. 1988. Bit error rate estimation for channels with memory. *IEEE Trans. Commun.*, COM-36(June):767–769.

Neyman, J. 1939. On a new class of contagious distribution, applicable in entomology and bacteriology. *Ann. Math. Stat.*, 10:35–57.

Rummler, W.D. 1979. A new selective fading model: Application to propagation data. *Bell Syst. Tech. J.*, 58(May/June):1037–1071.

Semmar, A., Lecours, M., Chouinard, J., and Ahern, J. 1991. Characterization of error sequences in UHF digital mobile radio channels. *IEEE Trans. Veh. Tech.*, 40(Nov.):769–776.

Smith, D.R. and Cormack, J.J. 1982. Measurement and characterization of a multipath fading channel. In *1982 International Conference on Communications*, pp. 7B.4.1–7B.4.6.

Stein, S., Schwartz, M., and Bennett, W. 1966. *Communication Systems and Techniques*, McGraw–Hill, New York.

Wozencraft, J.M. and Jacobs, I.M. 1967. *Principles of Communication Engineering*, Wiley, New York.

Further Information

A comprehensive treatment of channel models is given in *Fading Dispersive Communication Channels* by R.S. Kennedy. *Data Communications via Fading Channels*, an IEEE volume edited by K. Brayer, provides an excellent collection of papers on the subject.

Samsung Exhibit 1022, Page 42 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

14

Forward Error Correction Coding

14.1 Introduction 166
14.2 Fundamentals of Block Coding 167
14.3 Structure and Decoding of Block Codes 168
14.4 Important Classes of Block Codes 170
14.5 Principles of Convolutional Coding 171
14.6 Decoding of Convolutional Codes 174
14.7 Trellis-Coded Modulation 176
14.8 Additional Measures 177
14.9 Applications 177

V.K. Bhargava
University of Victoria

I.J. Fair
University of Victoria

14.1 Introduction

In 1948, Claude Shannon issued a challenge to communications engineers by proving that communication systems could be made arbitrarily reliable as long as a fixed percentage of the transmitted signal was redundant [Shannon, 1948]. He did not, however, indicate how this could be achieved. Subsequent research has led to a number of techniques that introduce redundancy to allow for correction of errors without retransmission. These techniques, collectively known as forward error correction (FEC) coding techniques, are used in systems where a reverse channel is not available for requesting retransmission, the delay with retransmission would be excessive, the expected number of errors would require a large number of retransmissions, or retransmission would be awkward to implement [Sklar, 1988].

A simplified model of a digital communication system which incorporates FEC coding is shown in Fig. 14.1. The FEC code acts on a **discrete data channel** comprising all system elements between the encoder output and decoder input. The encoder maps the source data to q-ary code symbols which are modulated and transmitted. During transmission, this signal can be corrupted, causing errors to arise in the demodulated symbol sequence. The FEC decoder attempts to correct these errors and restore the original source data.

A demodulator which outputs only a value for the q-ary symbol received during each symbol interval is said to make **hard decisions**. In the **binary symmetric channel** (BSC), hard decisions are made on binary symbols and the probability of error is independent of the value of the symbol. One example of a BSC is the coherently demodulated binary phase-shift-keyed (BPSK) signal corrupted by additive white Gaussian noise (AWGN). The conditional probability density functions which

0-8493-8349-8/97/$0.00+$.50
© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 43 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

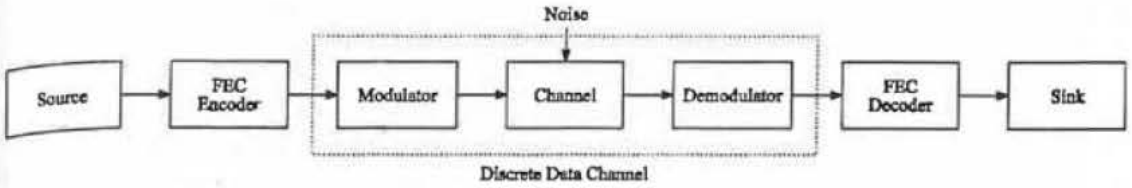


FIGURE 14.1 Block diagram of a digital communication system with forward error correction.

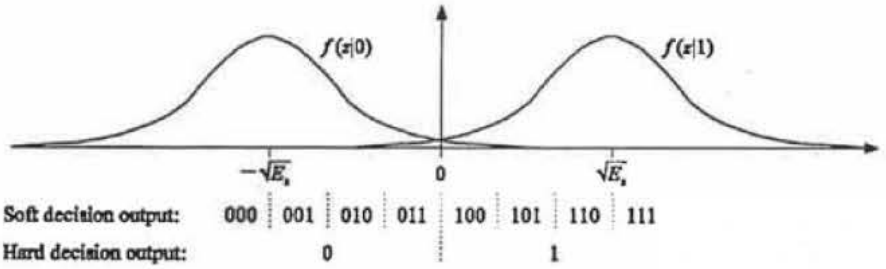


FIGURE 14.2 Hard and soft decision demodulation of a coherently demodulated BPSK signal corrupted by AWGN. $f(z \mid 1)$ and $f(z \mid 0)$ are the Gaussianly distributed conditional probability density functions at the threshold device.

result with this system are depicted in Fig. 14.2. The probability of error is given by the area under the density functions that lies across the decision threshold, and is a function of the symbol energy E_s and the one-sided noise power spectral density N_0.

Alternatively, the demodulator can make **soft decisions** or output an estimate of the symbol value along with an indication of its confidence in this estimate. For example, if the BPSK demodulator uses three-bit quantization, the two least significant bits can be taken as a confidence measure. Possible soft-decision thresholds for the BPSK signal are depicted in Fig. 14.2. In practice, there is little to be gained by using many soft-decision quantization levels.

Block and convolutional codes introduce redundancy by adding parity symbols to the message data. They map k source symbols to n code symbols and are said to have **code rate** $R = k/n$. With fixed information rates, this redundancy results in increased bandwidth and lower energy per transmitted symbol. At low signal-to-noise ratios, these codes cannot compensate for these impairments, and performance is degraded. At higher ratios of information symbol energy E_b to noise spectral density N_0, however, there is **coding gain** since the performance improvement offered by coding more than compensates for these impairments. Coding gain is usually defined as the reduction in required E_b/N_0 to achieve a specific error rate in an error-control coded system over one without coding. In contrast to block and convolutional codes, trellis-coded modulation introduces redundancy by expanding the size of the signal set rather than increasing the number of symbols transmitted, and so offers the advantages of coding to band-limited systems.

Each of these coding techniques is considered in turn. Following a discussion of interleaving and concatenated coding, this chapter concludes with a brief overview of FEC applications.

14.2 Fundamentals of Block Coding

In block codes there is a one-to-one mapping between k-symbol source words and n-symbol codewords. With q-ary signalling, q^k out of the q^n possible n-tuples are valid code vectors. The set of all n-tuples forms a **vector space** in which the q^k code vectors are distributed. The **Hamming**

Samsung Exhibit 1022, Page 44 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

distance between any two code vectors is the number of symbols in which they differ; the **minimum distance** $d_{\min}$ of the code is the smallest Hamming distance between any two codewords.

There are two contradictory objectives of block codes. The first is to distribute the code vectors in the vector space such that the distance between them is maximized. Then, if the decoder receives a corrupted vector, by evaluating the nearest valid code vector it will decode the correct word with high probability. The second is to pack the vector space with as many code vectors as possible to reduce the redundancy in transmission.

When code vectors differ in at least $d_{\min}$ positions, a decoder which evaluates the nearest code vector to each received word is guaranteed to correct up to t random symbol errors per word if

$$d_{\min} \geq 2t + 1 \tag{14.1}$$

Alternatively, all $q^n - q^k$ illegal words can be detected, including all error patterns with $d_{\min} - 1$ or fewer errors. In general, a block code can correct all patterns of t or fewer errors and detect all patterns of u or fewer errors provided that $u \geq t$ and

$$d_{\min} \geq t + u + 1 \tag{14.2}$$

If $q = 2$, knowledge of the positions of the errors is sufficient for their correction; if $q > 2$, the decoder must determine both the positions and values of the errors. If the demodulator indicates positions in which the symbol values are unreliable, the decoder can assume their value unknown and has only to solve for the value of these symbols. These positions are called **erasures**. A block code can correct up to t errors and v erasures in each word if

$$d_{\min} \geq 2t + v + 1 \tag{14.3}$$

14.3 Structure and Decoding of Block Codes

Shannon showed that the performance limit of codes with fixed code rate improves as the block length increases. As n and k increase, however, practical implementation requires that the mapping from message to code vector not be arbitrary but that an underlying structure to the code exist. The structures developed to date limit the error correcting capability of these codes to below what Shannon proved possible, on average, for a code with random codeword assignments. The search for good constructive codes continues.

A property which simplifies implementation of the coding operations is that of code linearity. A code is **linear** if the addition of any two code vectors forms another code vector, which implies that the code vectors form a subspace of the vector space of n-tuples. This subspace, which contains the all-zero vector, is spanned by any set of k linearly independent code vectors. Encoding can be described as the multiplication of the information k-tuple by a **generator matrix** $\boldsymbol{G}$, of dimension $k \times n$, which contains these basis vectors as rows. That is, a message vector $\boldsymbol{m}_i$ is mapped to a code vector $\boldsymbol{c}_i$ according to

$$\boldsymbol{c}_i = \boldsymbol{m}_i \boldsymbol{G}, \qquad i = 0, 1, \ldots, q^k - 1 \tag{14.4}$$

where elementwise arithmetic is defined in the **finite field** GF(q). In general, this encoding procedure results in code vectors with nonsystematic form in that the values of the message symbols cannot be determined by inspection of the code vector. However, if $\boldsymbol{G}$ has the form $[\boldsymbol{I}_k, \boldsymbol{P}]$ where $\boldsymbol{I}_k$ is the $k \times k$ identity matrix and $\boldsymbol{P}$ is a $k \times (n-k)$ matrix of parity checks, then the k most significant symbols of each code vector are identical to the message vector and the code has **systematic** form. This notation assumes that vectors are written with their most significant or first symbols in time on the left, a convention used throughout this chapter.

Samsung Exhibit 1022, Page 45 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

For each generator matrix there is an $(n-k) \times k$ **parity check matrix** $\boldsymbol{H}$ whose rows are orthogonal to the rows in $\boldsymbol{G}$, i.e., $\boldsymbol{G}\boldsymbol{H}^T = \boldsymbol{0}$. If the code is systematic, $\boldsymbol{H} = [-\boldsymbol{P}^T, \boldsymbol{I}_{n-k}]$. Since all codewords are linear sums of the rows in $\boldsymbol{G}$, it follows that $\boldsymbol{c}_i\boldsymbol{H}^T = \boldsymbol{0}$ for all $i, i = 0, 1, \ldots, q^k - 1$, and that the validity of the demodulated vectors can be checked by performing this multiplication. If a codeword $\boldsymbol{c}$ is corrupted during transmission so that the hard-decision demodulator outputs the vector $\hat{\boldsymbol{c}} = \boldsymbol{c} + \boldsymbol{e}$, where $\boldsymbol{e}$ is a nonzero error pattern, the result of this multiplication is an $(n-k)$-tuple that is indicative of the validity of the sequence. This result, called the **syndrome** $\boldsymbol{s}$, is dependent only on the error pattern since

$$\boldsymbol{s} = \hat{\boldsymbol{c}}\boldsymbol{H}^T = (\boldsymbol{c} + \boldsymbol{e})\boldsymbol{H}^T = \boldsymbol{c}\boldsymbol{H}^T + \boldsymbol{e}\boldsymbol{H}^T = \boldsymbol{e}\boldsymbol{H}^T \tag{14.5}$$

If the error pattern is a code vector, the errors go undetected. For all other error patterns, however, the syndrome is nonzero. Since there are $q^{n-k} - 1$ nonzero syndromes, $q^{n-k} - 1$ error patterns can be corrected. When these patterns include all those with t or fewer errors and no others, the code is said to be a **perfect code**. Few codes are perfect; most codes are capable of correcting some patterns with more than t errors. **Standard array decoders** use lookup tables to associate each syndrome with an error pattern but become impractical as the block length and number of parity symbols increases. Algebraic decoding algorithms have been developed for codes with stronger structure. These algorithms are simplified with imperfect codes if the patterns corrected are limited to those with t or fewer errors, a simplification called **bounded distance decoding**.

Cyclic codes are a subclass of linear block codes with an algebraic structure that enables encoding to be implemented with a linear feedback shift register and decoding to be implemented without a lookup table. As a result, most block codes in use today are cyclic or are closely related to cyclic codes. These codes are best described if vectors are interpreted as polynomials and the arithmetic follows the rules for polynomials where the elementwise operations are defined in GF(q). In a cyclic code, all codeword polynomials are multiples of a **generator polynomial** $g(x)$ of degree $n-k$. This polynomial is chosen to be a divisor of $x^n - 1$ so that a cyclic shift of a code vector yields another code vector, giving this class of codes its name. A message polynomial $m_i(x)$ can be mapped to a codeword polynomial $c_i(x)$ in nonsystematic form as

$$c_i(x) = m_i(x)g(x), \qquad i = 0, 1, \ldots, q^k - 1 \tag{14.6}$$

In systematic form, codeword polynomials have the form

$$c_i(x) = m_i(x)x^{n-k} - r_i(x), \qquad i = 0, 1, \ldots, q^k - 1 \tag{14.7}$$

where $r_i(x)$ is the remainder of $m_i(x)x^{n-k}$ divided by $g(x)$. Polynomial multiplication and division can be easily implemented with shift registers [Blahut, 1983].

The first step in decoding the demodulated word is to determine if the word is a multiple of $g(x)$. This is done by dividing it by $g(x)$ and examining the remainder. Since polynomial division is a linear operation, the resulting syndrome $s(x)$ depends only on the error pattern. If $s(x)$ is the all-zero polynomial, transmission is errorless or an undetectable error pattern has occurred. If $s(x)$ is nonzero, at least one error has occurred. This is the principle of the **cyclic redundancy check** (CRC). It remains to determine the most likely error pattern that could have generated this syndrome.

Single error correcting binary codes can use the syndrome to immediately locate the bit in error. More powerful codes use this information to determine the locations and values of multiple errors. The most prominent approach of doing so is with the iterative technique developed by Berlekamp. This technique, which involves computing an error-locator polynomial and solving for its roots, was subsequently interpreted by Massey in terms of the design of a minimum-length shift register. Once the location and values of the errors are known, Chien's search algorithm efficiently corrects

Samsung Exhibit 1022, Page 46 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

the code and is important for evaluating the performance of a code that uses threshold decoding. As i increases, $d_c(i)$ approaches the **free distance** of the code, d_{free}, which is the minimum Hamming distance in the set of arbitrarily long paths that diverge and then remerge in the trellis.

With maximum likelihood decoding, convolutional codes can generally correct up to t errors within three to five constraint lengths, depending on how the errors are distributed, where

$$d_{\text{free}} \geq 2t + 1 \tag{14.8}$$

The free distance can be calculated by exhaustively searching for the minimum-weight path that returns to the all-zero state, or evaluating the term of lowest degree in the generating function of the code.

The objective of a convolutional code is to maximize these distance properties. They generally improve as the constraint length of the code increases, and nonsystematic codes generally have better properties than systematic ones. Good codes have been found by computer search and are tabulated in many texts, including [Clark and Cain, 1981]. Convolutional codes with high code rate can be constructed by **puncturing** or periodically deleting coded symbols from a low rate code. A list of low rate codes and perforation matrices that result in good high rate codes can be found in many sources, including [Wu et al., 1987]. The performance of good punctured codes approaches that of the best convolutional codes known with similar rate, and decoder implementation is significantly less complex.

Convolutional codes can be **catastrophic**, having the potential to generate an unlimited number of decoded bit errors in response to a finite number of errors in the demodulated bit sequence. Catastrophic error propagation is avoided if the code has generator polynomials with a greatest common divisor of the form x^a for any a or, equivalently, if there are no closed-loop paths in the state diagram with all-zero output other than the one taken with all-zero input. Systematic codes are not catastrophic.

14.6 Decoding of Convolutional Codes

In 1967, Viterbi developed a maximum likelihood decoding algorithm that takes advantage of the trellis structure to reduce the complexity of the evaluation. This algorithm has become known as the **Viterbi algorithm**. With each received n-tuple, the decoder computes a **metric** or measure of likelihood for all paths that could have been taken during that interval and discards all but the most likely to terminate on each node. An arbitrary decision is made if path metrics are equal. The metrics can be formed using either hard or soft decision information with little difference in implementation complexity.

If the message has finite length and the encoder is subsequently flushed with zeros, a single decoded path remains. With a BSC, this path corresponds to the valid code sequence with minimum Hamming distance from the demodulated sequence. Full-length decoding becomes impractical as the length of the message sequence increases. The most likely paths tend to have a common stem, however, and selecting the trace value four or five times the constraint length prior to the present decoding depth results in near-optimum performance. Since the number of paths examined during each interval increases exponentially with the constraint length, the Viterbi algorithm also becomes impractical for codes with large constraint length. To date, Viterbi decoding has been implemented for codes with constraint lengths up to ten. Other decoding techniques, such as sequential and threshold decoding, can be used with larger constraint lengths.

Sequential decoding was proposed by Wozencraft, and the most widely used algorithm was developed by Fano. Rather than tracking multiple paths through the trellis, the sequential decoder operates on a single path while searching the code tree for a path with high probability. It makes tentative decisions regarding the transmitted sequence, computes a metric between its proposed path and the demodulated sequence, and moves forward through the tree as long as the metric indicates

Samsung Exhibit 1022, Page 47 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

ensures improved Euclidean distance is the method of **set partitioning**. This involves separating all parallel paths on the trellis with maximum distance and assigning the next greatest distance to paths that diverge from or merge onto the same node. Figures 14.6(c) and 14.6(d) give examples of mappings for the rate 2/3 code with 8-PSK and 8-PAM signal constellations respectively.

As with convolutional codes, the free distance of a TCM code is defined as the minimum distance between paths through the trellis, where the distance of concern is now Euclidean distance rather than Hamming distance. The free distance of an uncoded signal is defined as the distance between the closest signal points. When coded and uncoded signals have the same average power, the coding gain of the TCM system is defined as

$$\text{coding gain} = 20 \log_{10} \left(\frac{d_{\text{free, coded}}}{d_{\text{free, uncoded}}} \right) \tag{14.9}$$

It can be shown that the simple, rate 2/3 8 phase-shift keying (PSK) and 8 pulse-amplitude modulation (PAM) TCM systems provide gains of 3 dB and 3.3 dB, respectively [Clark and Cain, 1981]. More complex TCM systems yield gains up to 6 dB. Tables of good codes are given in [Ungerboeck, 1987].

14.8 Additional Measures

When the demodulated sequence contains bursts of errors, the performance of codes designed to correct independent errors improves if coded sequences are **interleaved** prior to transmission and deinterleaved prior to decoding. Deinterleaving separates the burst errors, making them appear more random and increasing the likelihood of accurate decoding. It is generally sufficient to interleave several block lengths of a block coded signal or several constraint lengths of a convolutionally encoded signal. Block interleaving is the most straightforward approach, but delay and memory requirements are halved with convolutional and helical interleaving techniques. Periodicity in the way sequences are combined is avoided with pseudorandom interleaving.

Concatenated codes, first investigated by Forney, use two levels of coding to achieve a level of performance with less complexity than a single coding stage would require. The inner code interfaces with the modulator and demodulator and corrects the majority of the errors; the outer code corrects errors that appear at the output of the inner-code decoder. A convolutional code with Viterbi decoding is usually chosen as the inner code, and an RS code is often chosen as the outer code due to its ability to correct the bursts of bit errors which can result with incorrect decoding of trellis-coded sequences. Interleaving and deinterleaving outer-code symbols between coding stages offers further protection against the burst error output of the inner code.

14.9 Applications

FEC coding remained of theoretical interest until advances in digital technology and improvements in decoding algorithms made their implementation possible. It has since become an attractive alternative to improving other system components or boosting transmission power. FEC codes are commonly used in digital storage systems, deep-space and satellite communication systems, terrestrial radio and band limited wireline systems, and have also been proposed for fiber optic transmission. Accordingly, the theory and practice of error correcting codes now occupies a prominent position in the field of communications engineering.

Deep-space systems began using forward error correction in the early 1970s to reduce transmission power requirements, and used multiple error correcting RS codes for the first time in 1977 to protect against corruption of compressed image data in the Voyager missions [Wicker and Bhargava, 1994]. The Consultative Committee for Space Data Systems (CCSDS) has since recommended use of a concatenated coding system which uses a rate 1/2, constraint length 7 convolutional inner code and a (255, 223) RS outer code.

Samsung Exhibit 1022, Page 48 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Coding is now commonly used in satellite systems to reduce power requirements and overall hardware costs and to allow closer orbital spacing of geosynchronous satellites [Berlekamp et al., 1987]. FEC codes play integral roles in the VSAT, MSAT, INTELSAT, and INMARSAT systems [Wu et al., 1987]. Further, a (31, 15) RS code is used in the joint tactical information distribution system (JTIDS), a (7, 2) RS code is used in the air force satellite communication system (AFSATCOM), and a (204, 192) RS code has been designed specifically for satellite time division multiple access (TDMA) systems. Another code designed for military applications involves concatenation of a Golay and RS code with interleaving to ensure an imbalance of 1's and 0's in the transmitted symbol sequence and enhance signal recovery under severe noise and interference [Berlekamp et al., 1987].

TCM has become commonplace in transmission of data over voiceband telephone channels. Modems developed since 1984 use trellis coded QAM modulation to provide robust communication at rates above 9.6 kb/s.

FEC codes have also been widely used in digital recording systems, most prominently in the compact disc digital audio system. This system uses two levels of coding and interleaving in the cross-interleaved RS coding (CIRC) system to correct errors that result from disc imperfections and dirt and scratches which accumulate during use. Steps are also taken to mute uncorrectable sequences [Wicker and Bhargava, 1994].

Defining Terms

Binary symmetric channel: A memoryless discrete data channel with binary signalling, hard-decision demodulation, and channel impairments that do not depend on the value of the symbol transmitted.

Bounded distance decoding: Limiting the error patterns which are corrected in an imperfect code to those with t or fewer errors.

Catastrophic code: A convolutional code in which a finite number of code symbol errors can cause an unlimited number of decoded bit errors.

Code rate: The ratio of source word length to codeword length, indicative of the amount of information transmitted per encoded symbol.

Coding gain: The reduction in signal-to-noise ratio required for specified error performance in a block or convolutional coded system over an uncoded system with the same information rate, channel impairments, and modulation and demodulation techniques. In TCM, the ratio of the squared free distance in the coded system to that of the uncoded system.

Column distance: The minimum Hamming distance between convolutionally encoded sequences of a specified length with different leading n-tuples.

Concatenated codes: Two levels of codes that achieve a level of performance with less complexity than a single coding stage would require. The inner code is often a convolutional code, and the outer code is usually an RS code.

Cyclic code: A block code in which cyclic shifts of code vectors are also code vectors.

Cyclic redundancy check: When the syndrome of a cyclic block code is used to detect errors.

Designed distance: The guaranteed minimum distance of a BCH code designed to correct up to t errors.

Discrete data channel: The concatenation of all system elements between FEC encoder output and decoder input.

Distance profile: The minimum Hamming distance after each encoding interval of convolutionally encoded sequences which differ in the first interval.

Erasure: A position in the demodulated sequence where the symbol value is unknown.

Finite field: A finite set of elements and operations of addition and multiplication that satisfy specific properties. Often called Galois fields and denoted GF(q), where q is the number of elements in the field. Finite fields exist for all q which are prime or the power of a prime.

Samsung Exhibit 1022, Page 49 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Free distance: The minimum Hamming weight of convolutionally encoded sequences that diverge and remerge in the trellis. Equals the maximum column distance and the limiting value of the distance profile.

Generator matrix: A matrix used to describe a linear code. Code vectors equal the information vectors multiplied by this matrix.

Generator polynomial: The polynomial that is a divisor of all codeword polynomials in a cyclic block code; a polynomial that describes circuit connections in a convolutional encoder.

Hamming distance: The number of symbols in which codewords differ.

Hard decision: Demodulation that outputs only a value for each received symbol.

Interleaving: Shuffling the coded bit sequence prior to modulation and reversing this operation following demodulation. Used to separate and redistribute burst errors over several codewords (block codes) or constraint lengths (trellis codes) for higher probability of correct decoding by codes designed to correct random errors.

Linear code: A code whose code vectors form a vector space. Equivalently, a code where the addition of any two code vectors forms another code vector.

Maximum distance separable: A code with the largest possible minimum distance given the block length and code rate. These codes meet the Singleton bound of $d_{\min} \leq n - k + 1$.

Metric: A measure of goodness against which items are judged. In the Viterbi algorithm, an indication of the probability of a path being taken given the demodulated symbol sequence.

Minimum distance: In a block code, the smallest Hamming distance between any two codewords. In a convolutional code, the column distance after K intervals.

Parity check matrix: A matrix whose rows are orthogonal to the rows in the generator matrix of a linear code. Errors can be detected by multiplying the received vector by this matrix.

Perfect code: A t error correcting (n, k) block code in which $q^{n-k} - 1 = \sum_{i=1}^{t} \binom{n}{i}$.

Puncturing: Periodic deletion of code symbols from the sequence generated by a convolutional encoder for purposes of constructing a higher rate code. Also, deletion of parity bits in a block code.

Set partitioning: Rules for mapping coded sequences to points in the signal constellation that always result in a larger Euclidean distance for a TCM system than an uncoded system, given appropriate construction of the trellis.

Soft decision: Demodulation that outputs an estimate of the received symbol value along with an indication of the reliability of this value. Usually implemented by quantizing the received signal to more levels than there are symbol values.

Standard array decoding: Association of an error pattern with each syndrome by way of a lookup table.

Syndrome: An indication of whether or not errors are present in the demodulated symbol sequence.

Systematic code: A code in which the values of the message symbols can be identified by inspection of the code vector.

Vector space: An algebraic structure comprised of a set of elements in which operations of vector addition and scalar multiplication are defined. For our purposes, a set of n-tuples consisting of symbols from GF(q) with addition and multiplication defined in terms of elementwise operations from this finite field.

Viterbi algorithm: A maximum-likelihood decoding algorithm for trellis codes that discards low-probability paths at each stage of the trellis, thereby reducing the total number of paths that must be considered.

References

Berlekamp, E.R., Peile, R.E., and Pope, S.P. 1987. The application of error control to communications. *IEEE Commun. Mag.* 25(4):44–57.

Samsung Exhibit 1022, Page 50 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

15

Automatic Repeat Request

15.1 Introduction 181
15.2 Fundamentals and Basic Automatic Repeat Request Schemes . 182
Basic Principles • Stop-and-Wait Automatic Repeat Request • Sliding Window Protocols
15.3 Performance Analysis and Limitations 187
Stop-and-Wait Automatic Repeat Request • Continuous Automatic Repeat Request: Sliding Window Protocols
15.4 Variants of the Basic Automatic Repeat Request Schemes 193
15.5 Hybrid Forward Error Control/Automatic Repeat Request Schemes 194
15.6 Application Problem 195
Solution
15.7 Conclusion 197

David Haccoun
École Polytechnique de Montréal

Samuel Pierre
Université du Québec

15.1 Introduction

In most digital communication systems, whenever error events occur in the transmitted messages, some action must be taken to correct these events. This action may take the form of an **error correction** procedure that attempts to correct these errors. In some applications where a two-way communication link exists between the sender and the receiver, the receiver may inform the sender that a message has been received in error and, hence, request a repeat of that message. In principle, the procedure may be repeated as many times as necessary until that message is received error free. An error control system in which the erroneously received messages are simply retransmitted is called *automatic-repeat-request* (ARQ).

In ARQ systems, the receiver must perform only an **error detection** procedure on the received messages without attempting to correct the errors. Hence, an error detecting code, in the form of specific redundant or parity-check symbols, must be added to the information-bearing sequence. In general, as the error detecting capability of the code increases, the number of added redundant symbols must also be increased. Clearly, with such a system, an erroneously received message is delivered to the user only if the receiver fails to detect the presence of errors. Since error detection coding is simple, powerful, and quite robust, ARQ systems constitute a simple and efficient method for providing highly reliable transfer of messages from the source to the user over a variety of transmission channels. ARQ systems are therefore widely used in data communication systems that are highly sensitive to errors, such as in computer to computer communications.

0-8493-8349-8/97/$0.00+$.50

© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 51 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

This chapter presents and discusses principles, performances, limitations, and variants of basic ARQ strategies. The fundamentals of the basic ARQ schemes are presented in Sec. 15.2, whereas the performance analysis and the limitations are carried out in Sec. 15.3. Section 15.4 presents some of the most common variants of the basic ARQ schemes, and hybrid forward error control (FEC)/ARQ techniques are outlined in Sec. 15.5. Finally, an application problem is provided in Sec. 15.6.

15.2 Fundamentals and Basic Automatic Repeat Request Schemes

Messages transmitted on a communication channel are subjected to errors, which must be detected and/or corrected at the receiver. This section presents the basic principles underlying the concept of error control, using retransmission schemes.

Basic Principles

In the seven-layer open system interconnection (OSI) model, error control refers to some basic functions that may be performed at several layers, especially at the *transport layer* (level 4) and at the *data link layer* (level 2). The transport layer is responsible for the end-to-end transport of information processed as *protocol data unit* (PDU) through a network infrastructure. An intermediate layer, called *network layer* (level 3), ensures that messages, which are broken down into smaller blocks called *packets*, can successfully be switched from one node to another of the network.

The error control on each individual link of the network is performed by the data link layer, which transforms packets received from the network layer into *frames* by adding address, control, and error check fields prior to their transmission into the channel. Figure 15.1 shows an example of the frame structure, whose length, which depends on the particular data link protocol used, usually varies approximately between 50 and 200 bytes. The length of each field also depends on the particular data link control protocol employed. The data link layer must ensure a correct and orderly delivery of packets between switching nodes in the network. The objective of the error control is to maintain the integrity of the data in the frames as they transit through the links, overcoming the loss, duplication, desequencing, and damage of data and control information bits at these various levels. Most ARQ error control techniques integrate error detection using standard *cyclic redundancy check* (CRC), as well as requests for retransmission using PDU.

Error control coding consists essentially of adding to the information bits to be transmitted some extra or redundant bits, called *parity-check* bits, on a regular and controlled manner. These parity-check bits, which are removed at the receiver, do not convey information but allow the receiver to detect and possibly correct errors in the received frames.

The single parity-check bit used in ASCII is an example of a redundant bit for error detection purposes. It is an eighth bit added to the seven information bits in the frame representing the alphanumeric character to be transmitted. This bit takes the value 1 if the number of 1s in the first

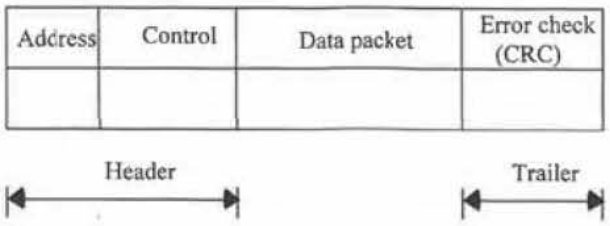


FIGURE 15.1 Structure of a frame.

Samsung Exhibit 1022, Page 52 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

seven bits is odd (odd parity), and 0 otherwise. At the receiver, the parity check bit of the received frame is computed again in order to verify whether or not it agrees with the received parity check bit. A mismatch of these parity check bits indicates that an odd number of errors has occurred during transmission, leading the receiver to request a retransmission of the erroneous frame.

ARQ refers to retransmission techniques, which basically operate as follows. Erroneously received frames are retransmitted until they are received or detected as being error free; errors are detected using a simple detection code. Positive (ACK) or negative (NAK) acknowledgments are sent back by the receiver to the sender over a reliable **feedback channel**, in order to report whether a previously transmitted frame has been received error free or with errors. In principle, a positive acknowledgment signals the transmitter to send the next data frame, whereas a negative acknowledgment is a request for frame retransmission. *stop-and-wait, go-back-N,* and *selective-repeat* are the three basic versions of ARQ, the last two schemes being usually referred to as *sliding window protocols.*

Stop-and-Wait Automatic Repeat Request

Stop-and-wait ARQ is a very simple acknowledgment scheme, which works as follows. Sender and receiver communicate through a **half-duplex** point-to-point **link**. Following transmission of a single frame to the receiver, the sender waits for an acknowledgment from the receiver before sending the next data frame or repeating the same frame. Upon receiving a frame, the receiver sends back to the sender a positive acknowledgment (ACK) if the frame is correct; otherwise it transmits a negative acknowledgment (NAK).

Figure 15.2 illustrates the frame flow for stop-and-wait ARQ. Frames $F1$, $F2$, $F3$, ... are in queue for transmission. At time t_0, the sender transmits frame $F1$ to the receiver and stays idle until either an ACK or NAK is received for that frame. At time t_1, $F1$ is correctly received, hence an ACK is sent back. At time t_2, the sender receives this ACK and transmits the next frame $F2$, which is received at time t_3 and detected in error. Frame $F2$ is then discarded at the receiver and a NAK is therefore returned to the sender. At time t_4, the sender receiving a NAK for $F2$ retransmits a copy of that frame and waits again, until it receives at time t_6 either an ACK or NAK for $F2$. Assuming it is an ACK, as illustrated in Fig. 15.2, the sender then proceeds to transmit the next frame $F3$, and so on. Clearly, such a technique requires the sender to store in a buffer a copy of a transmitted frame until a positive acknowledgment has been received for that frame.

Obviously, this technique does not take into account all contingencies. Indeed, it is possible that some transmitted frames damaged by transmission impairments do not even reach the receiver, in which case the receiver cannot even acknowledge them. To overcome this situation, a timer is integrated in the sender for implementing a *time-out* mechanism. A time out is some pre-established

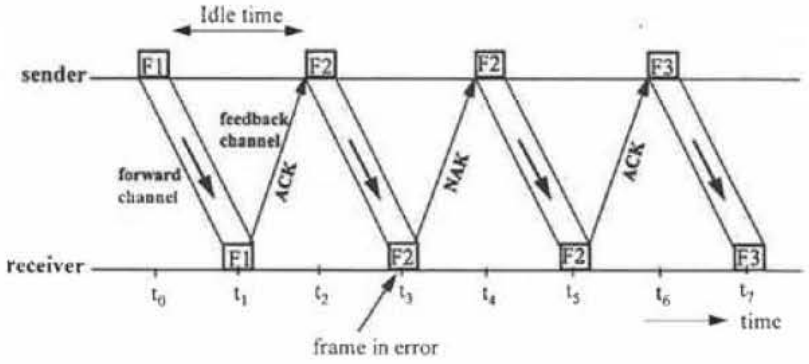


FIGURE 15.2 Stop-and-wait ARQ.

Samsung Exhibit 1022, Page 53 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

time interval beyond which a transmitted frame is considered lost if its acknowledgment is not received; a timed-out frame is usually simply retransmitted.

Another undesirable situation can be described as follows. The sender transmits a copy of some frame F, which is received correctly by the receiver. A positive acknowledgment (ACK) is then returned to the sender. Unfortunately, this ACK may be corrupted during its transmission in the feedback channel so that it becomes unrecognizable by the sender which therefore transmits another copy of the same frame. As a result, there is a possibility of frame duplication at the receiver leading to an ambiguity between a frame and its immediate predecessor or successor. This ambiguity can be easily resolved by attaching a sequence number in each transmitted frame. In practice, a single-bit (0 or 1) inserted in the header of each frame is sufficient to distinguish a frame from its duplicate at the receiver.

Stop-and-wait ARQ schemes offer the advantage of a great simplicity. However, they are not very suitable for high-speed modern communication systems since data is transmitted in one direction only. Protocols in which the ACK for a frame must be received by the sender prior to the transmission of the next frame are referred to as *positive acknowledgment with retransmission* (PAR) protocols. The principal weakness of PAR schemes is an inefficient link utilization related to the time lost awaiting for acknowledgments before sending another frame, especially if the propagation delay is significantly longer than the packet transmission time, such as in satellite links. To overcome this shortcoming, PAR schemes have been adapted to more practical situations, which require transmitting frames in both directions, using **full-duplex links** between sender and receiver. Clearly then, each direction of transmission acts as the **forward channel** for the frames and as the return channel for the acknowledgments. Now in order to improve further the efficiency of each channel, ACK/NAK control packets and information frames need not be sent separately. Instead, the acknowledgment about each previously received frame is simply appended to the next transmitted frame. In this technique called *piggybacking*, clearly, transmission of the acknowledgment of a received frame is delayed by the transmission of the next data frame. A fixed maximum waiting time for the arrival of a new frame is thus set. If no new frame has arrived by the end of that waiting time, then the acknowledgment frame is sent separately. Even with piggybacking, PAR schemes still require the sender to wait for an acknowledgment before transmitting another frame. An obvious improvement of the efficiency may be achieved by relaxing this condition and allow the sender to transmit up to W, $W > 1$ frames without waiting for an acknowledgment. A further improvement still is to allow a continuous transmission of the frames without any waiting for acknowledgments. These protocols are called *Continuous ARQ* and are associated with the concept of sliding window protocols.

Sliding Window Protocols

To overcome the inefficiency of PAR schemes, an obvious solution is not to stay idle between frame transmissions. One approach is to use a sliding window protocol where a window refers to a subset of consecutive frames.

Consider a window that contains N frames numbered $\omega, \omega + 1, \ldots, \omega + N - 1$, as shown in Fig. 15.3. Every frame using a number smaller than ω has been sent and acknowledged as identified in Fig. 15.3 as *frames before the window*, whereas no frame with a number larger than or equal to $\omega + N$, identified as *frames beyond the window*, has yet been sent. Frames in the window that have been sent but that have not yet been acknowledged are said to be *outstanding frames*. As outstanding frames are being acknowledged, the window shifts to the right to exclude acknowledged frames and, subsequently, to include the new frames that have to be sent. Since the window must always contain frames that are numbered consecutively, the frames are excluded from the window in the same order in which they were included.

To limit the number of outstanding frames, a limit on the window size may be determined. When this limit is reached, the sender accepts no more frames. A window size $N = 1$ corresponds to the

Samsung Exhibit 1022, Page 54 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

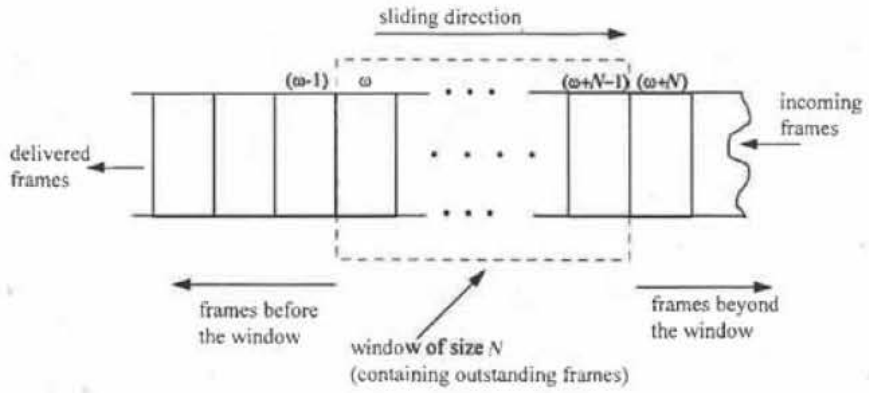


FIGURE 15.3 Frame sequencing in a sliding window protocol (sender side).

stop-and-wait protocol, whereas a window size greater than the total number of frames that can be represented in the header field of the frame refers to an *unrestricted protocol.* Clearly, the window size has some influence on the network traffic and on buffering requirements. Go-back-N and selective-repeat may be considered as two common implementations of a sliding window protocol as will be described next.

Go-Back-N Automatic Repeat Request

Go-back-N ARQ schemes use continuous transmission without waiting for ACK between frames. Clearly, a full-duplex link between the sender and the receiver is required, allowing a number of consecutive frames to be sent without receiving an acknowledgment. In fact, ACKs may not even be transmitted. Upon detecting a frame in error, the receiver sends a NAK for that frame and discards that frame and all succeeding frames, until that erroneous frame has been correctly received. Following reception of a NAK, the sender completes the transmission of the current frame and retransmits the erroneous frame and all subsequent ones.

In accordance with the sliding window protocol, each frame is numbered by an integer $\ell = 0, 1, 2, \ldots, 2^k - 1$, where k denotes the number of bits in the number field of the frame. In general, in order to facilitate their interpretation, frames are all numbered consecutively modulo-2^k. For example, for $k = 3$, frames are numbered 0–7 repeatedly.

It is important to fix the maximum window size, that is, the maximum number of outstanding frames waiting to be transmitted at any one time. To avoid two different outstanding frames having the same number, as well as to prevent pathological protocol failures, the maximum window size is $N = 2^k - 1$. It follows that the number of discarded frames is at most equal to N; however, the actual exact number of discarded frames depends also on the propagation delay. For most usual applications, a window of size $N = 7$ (i.e., $k = 3$) is adequate, whereas for some satellite links the window size N is usually equal to 127.

Figure 15.4 illustrates the frame flow for a go-back-N ARQ scheme, with $k = 3$ and, hence, $N = 7$. Thus the sender can transmit sequences of frames numbered $F0, F1, F2, \ldots, F7$. As shown in Fig. 15.4, since the receiver detects an error on frame $F3$, it returns NAK3 to the sender and discards both $F3$ and the succeeding frames $F4$, $F5$, and $F6$, which because of propagation delays have already been transmitted by the sender before it received the negative acknowledgment NAK3. Then, the sender retransmits $F3$ and the sequence of frames $F4$, $F5$, and $F6$, and proceeds to transmit $F7$ (using the current sequencing) as well as frames $F0$, $F1$, ..., using the next sequencing. Obviously should any of the preceding frames be received in error, then the same repeat procedure is implemented, starting from that frame in error.

Go-back-N procedures have been implemented in various data link protocols such as high-level data link control (HDLC), which refers to layer 2 of the seven-layer OSI model, and transmission

Samsung Exhibit 1022, Page 55 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

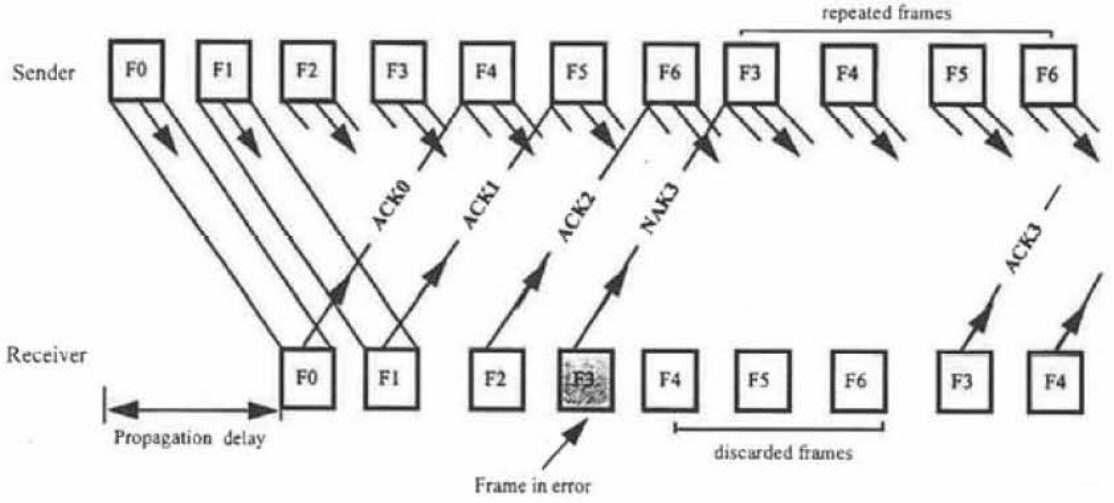


FIGURE 15.4 Go-back-N ARQ.

control protocol/internet protocol (TCP/IP), which supports the Internet, the world's largest interconnections of computer networks. Indeed, to ensure reliable end-to-end connections and improve the performance, TCP uses a sliding window protocol with timers. In accordance with this protocol, the sender can transmit several TCP messages or frames before an acknowledgment is received for any of the frames. The number of frames allowed in transit is negotiated dynamically using the window field in the TCP header.

As an implementation of the sliding window protocol, the go-back-N scheme is more efficient and more general than the stop-and-wait scheme, which is clearly equivalent to go-back-N, with $N = 1$. However, go-back-N presents the disadvantage of requiring a prerequisite numbering of frames as well as the buffering in the window of N frames awaiting positive acknowledgments. Furthermore, in the absence of a controlled source, all incoming frames not yet accepted by the window must be buffered while they wait to be inserted in the window. Inefficiency is also due to the discarding of all frames that follow the frame received in error, even though these frames may be error free. On the other hand, such a procedure presents the advantage of maintaining the proper sequencing of the frames accepted at the receiver.

Selective-Repeat Automatic Repeat Request

To overcome the inefficiency resulting from unnecessarily retransmitting the error-free frames, retransmission can be restricted to only those frames that have been detected in error. Repeated frames correspond to erroneous or damaged frames that are negatively acknowledged as well as frames for which the time out has expired. Such a procedure is referred to as selective-repeat ARQ, which obviously should improve the performance over both stop-and-wait and go-back-N schemes. Naturally, a full-duplex link is again assumed.

Figure 15.5 illustrates the frame flow for selective-repeat ARQ. The frames appear to be disjointed but there is no idle time between consecutive frames. The sender transmits a sequence of frames $F1, F2, F3, \ldots$. As illustrated in Fig. 15.5, the receiver detects an error on the third frame $F3$ and thus returns a negative acknowledgment denoted NAK3 to the sender. However, due to the continuous nature of the protocol, the succeeding frames $F4$, $F5$, and $F6$ have already been sent and are in the pipeline by the time the sender receives NAK3. Upon receiving NAK3, the sender completes the transmission of the current frame ($F6$) and then retransmits $F3$ before sending the next frame $F7$ in the sequence. The correctly received but incomplete sequence $F4$, $F5$, $F6$ must

Samsung Exhibit 1022, Page 56 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

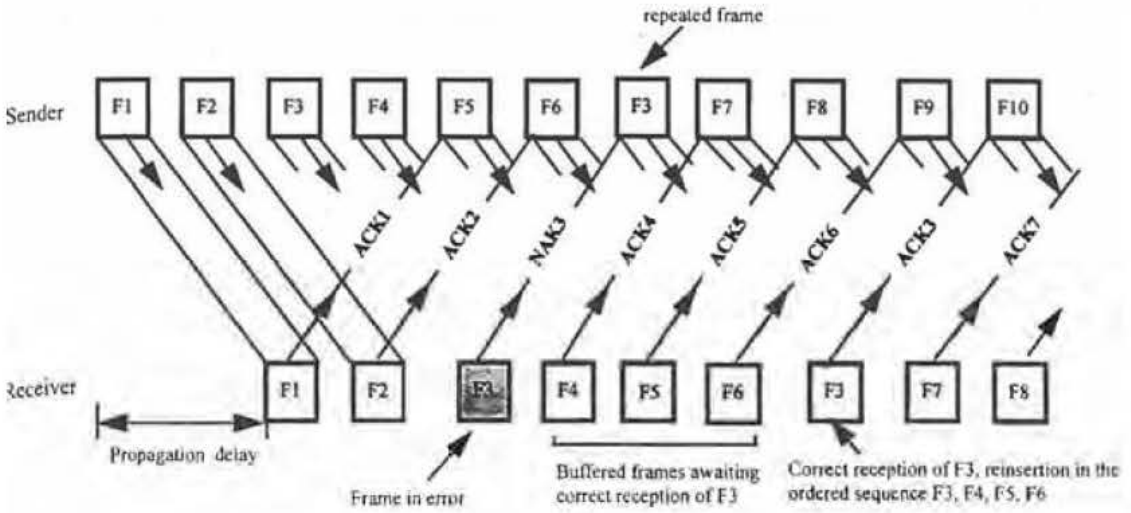


FIGURE 15.5 Selective-repeat ARQ.

be buffered at the receiver, until $F3$ has been correctly received and inserted in its proper place to complete the sequence $F3, F4, F5, F6, \ldots$, which is then delivered.

If $F3$ is correctly received at the first retransmission, the buffered sequence is $F4$, $F5$, and $F6$. However, if $F3$ is received in error during its retransmission and hence must be retransmitted again, then clearly the buffered sequence is $F4, F5, F6, F7, F8, F9$. Therefore, multiple retransmissions of a given frame leads to an ever larger buffering, which may be theoretically unbounded. To circumvent such a possibility and the ensuing buffer overflow, a time out mechanism is usually incorporated in all practical selective-repeat ARQ systems.

In practice, each correctly received frame prior to the appearance of an erroneous frame denoted F is delivered to the user (i.e., the physical layer). Thereafter, all error free successive frames are buffered at the receiver, until frame F is correctly received. It is then inserted at the beginning of the buffered sequence, which can thus be safely delivered to the user up to the next frame received in error. However, the reordering becomes somewhat more complex if other frames are received in error before frame F is finally received error free. As a consequence, proper buffering must be provided by the sender, which must also integrate the appropriate logic in order to send frames that are out of sequence. The receiver must also be provided with adequate buffering to store the out-of-order frames within the window until the frame in error is correctly received, in addition to be provided with the appropriate logic for inserting the accepted frames in their proper place in the sequence. The required buffers may be rather large for satellite transmission with long propagation delays and long pipelines corresponding to the many frames in transit between sender and receiver. As a result, selective-repeat ARQ procedure tends to be less commonly used than the go-back-N ARQ scheme, even though it may be more efficient from a throughput point of view.

15.3 Performance Analysis and Limitations

In ARQ systems, the performance is usually measured using two parameters: throughput efficiency and undetected error probability on the data bits. The throughput efficiency is defined as the ratio of the average number of information bits per second delivered to the user, to the average number of bits per second that have been transmitted in the system. This throughput efficiency is obviously smaller than 100%. For example, using an error-detecting scheme with a coding rate $R = 0.98$, an error-free transmission would then correspond to a throughput efficiency of 98%, and clearly, any frame retransmission yields a further decrease of the throughput efficiency.

Samsung Exhibit 1022, Page 57 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

that they are identical inverses of each other). That is, after **despreading** the received signal would be identical to the transmitted signal prior to **spreading**. In the presence of narrowband interference, however, there is a significant advantage to employing the spreading/despreading procedure described. The reason for this is as follows. Since the interference is introduced after the transmitted signal is spread, then, whereas the despreading operation at the receiver shrinks the desired signal back to its original bandwidth, at the same time it spreads the undesired signal (interference) in bandwidth by the same amount, thus reducing its power spectral density. This, in turn, serves to diminish the effect of the interference on the receiver performance, which depends on the amount of interference power in the despread bandwidth. It is indeed this very simple explanation, which is at the heart of all spread spectrum techniques.

16.3 Basic Concepts and Terminology

To describe this process analytically and at the same time introduce some terminology that is common in spread spectrum parlance, we proceed as follows. Consider a communicator that desires to send a message using a transmitted power S Watts (W) at an information rate R_b bits/s (bps). By introducing a SS modulation, the bandwidth of the transmitted signal is increased from R_b Hz to W_{ss} Hz where $W_{ss} \gg R_b$ denotes the **spread spectrum bandwidth**. Assume that the channel introduces, in addition to the usual thermal noise (assumed to have a single-sided power spectral density (PSD) equal to N_0 W/Hz), an additive interference (jamming) having power J distributed over some bandwidth W_J. After despreading, the desired signal bandwidth is once again now equal to R_b Hz and the interference PSD is now $N_J = J/W_{ss}$. Note that since the thermal noise is assumed to be white, i.e., it is uniformly distributed over all frequencies, its PSD is unchanged by the despreading operation and, thus, remains equal to N_0. Regardless of the signal and interferer waveforms, the equivalent bit energy-to-total noise spectral density ratio is, in terms of the given parameters,

$$\frac{E_b}{N_t} = \frac{E_b}{N_0 + N_J} = \frac{S/R_b}{N_0 + J/W_{ss}} \tag{16.1}$$

For most practical scenarios, the jammer limits performance and, thus, the effects of receiver noise in the channel can be ignored. Thus, assuming $N_J \gg N_0$, we can rewrite Eq. (16.1) as

$$\frac{E_b}{N_t} \simeq \frac{E_b}{N_J} = \frac{S/R_b}{J/W_{ss}} = \frac{S}{J}\frac{W_{ss}}{R_b} \tag{16.2}$$

where the ratio J/S is the *jammer-to-signal power ratio* and the ratio W_{ss}/R_b is the **spreading ratio** and is defined as the **processing gain** of the system. Since the ultimate error probability performance of the communication receiver depends on the ratio E_b/N_J, we see that from the communicator's viewpoint his goal should be to minimize J/S (by choice of S) and maximize the processing gain (by choice of W_{ss} for a given desired information rate). The possible strategies for the jammer will be discussed in the section on military applications dealing with AJ communications.

16.4 Spread Spectrum Techniques

By far the two most popular spreading techniques are **direct sequence (DS) modulation** and **frequency hopping (FH) modulation**. In the following subsections, we present a brief description of each.

Samsung Exhibit 1022, Page 58 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

21

Digital Modulation Techniques

21.1 Introduction 273
21.2 The Challenge of Digital Modulation 274
Bandwidth • Signal-to-Noise Ratio • Error Probability
21.3 One-Dimensional Modulation: Pulse-Amplitude Modulation (PAM) 278
21.4 Two-Dimensional Modulations 279
Phase-Shift Keying (PSK) • Quadrature Amplitude Modulation (QAM)
21.5 Multidimensional Modulations: Frequency-Shift Keying (FSK) 281
21.6 Multidimensional Modulations: Lattices 282
Examples of Lattices • The Coding Gain of a Lattice • Carving a Signal Constellation out of a Lattice
21.7 Modulations with Memory 284

Ezio Biglieri
Politecnico di Torino

21.1 Introduction

The goal of a digital communication system is to deliver information represented by a sequence of binary symbols, through a physical channel, to a user. The mapping of these symbols into signals, selected to match the features of the physical channel, is called **digital modulation**.

The digital modulator is the device used to achieve this mapping. The simplest type of modulator has no memory, that is, the mapping of blocks of binary digits into signals is performed independently of the blocks transmitted before or after. If the modulator maps the binary digits 0 and 1 into a set of two different waveforms, then the modulation is called *binary*. Alternatively, the modulator may map symbols formed by h binary digits at a time onto $M = 2^h$ different waveforms. This modulation is called *multilevel*.

The general expression of a signal modulated by a modulator with memory is

$$v(t) = \sum_{n=-\infty}^{\infty} s(t - nT; \xi_n, \sigma_n)$$

where $\{s(t; i, j)\}$ is a set of waveforms, ξ_n is the symbol emitted by the source at time nT, and σ_n is the state of the modulator at time n. If the modulator has no memory, then there is no dependence on σ_n. A *linear* memoryless modulation scheme is one such that $s(t - nT; \xi_n; \sigma_n) = \xi_n s(t - nT)$.

0-8493-8349-8/97/$0.00+$.50

© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 59 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Given a signal set $\{s_i(t)\}_{i=1}^{M}$ used for modulation, its compact characterization may be given in terms of a **geometric representation**. From the set of M waveforms we first construct a set of $N \leq M$ orthonormal waveforms $\{\psi_i(t)\}_{i=1}^{N}$ (the Gram–Schmidt procedure is the appropriate algorithm to do this). Next we express each signal $s_k(t)$ as a linear combination of these waveforms,

$$s_k(t) = \sum_{i=1}^{N} s_{ki}\psi_i(t)$$

The coefficients of this expansion may be interpreted as the components of M vectors that geometrically represent the original signal set, or, equivalently, as the coordinates in a Euclidean N-dimensional space of a set of points called the **signal constellation**.

21.2 The Challenge of Digital Modulation

The selection of a digital modulation scheme should be done by making the best possible use of the resources available for transmission, namely, bandwidth, power, and complexity, in order to achieve the reliability required.

Bandwidth

There is no unique definition of signal bandwidth. Actually, any signal $s(t)$ strictly limited to a time interval T would have an infinite bandwidth if the latter were defined as the support of the Fourier transform of $s(t)$. For example, consider the bandpass linearly modulated signal

$$v(t) = \Re\left[\sum_{n=-\infty}^{\infty} \xi_k s(t - nT)e^{j2\pi f_0 t}\right]$$

where $\Re$ denotes real part, f_0 is the carrier frequency, $s(t)$ is a rectangular pulse with duration T and amplitude 1, and (ξ_k) is a stationary sequence of complex uncorrelated random variables with $\mathrm{E}(\xi_n) = 0$ and $\mathrm{E}(|\xi_n|^2) = 1$. Then the power density spectrum of $v(t)$ is given by

$$\mathcal{G}(f) = \frac{1}{4}\left[G(-f - f_0) + G(f - f_0)\right]$$

where

$$G(f) = T\left[\frac{\sin \pi f T}{\pi f T}\right]^2 \tag{21.1}$$

The function $G(f)$ is shown in Fig. 21.1.

The following are possible definitions of the bandwidth:

- *Half-power bandwidth.* This is the interval between the two frequencies at which the power spectrum is 3 dB below its peak value.
- *Equivalent noise bandwidth.* This is given by

$$B_{\mathrm{eq}} = \frac{1}{2}\frac{\int_{-\infty}^{\infty} \mathcal{G}(f)\,\mathrm{d}f}{\max_f \mathcal{G}(f)}$$

This measures the basis of a rectangle whose height is $\max_f \mathcal{G}(f)$ and whose area is one-half of the power of the modulated signal.

Samsung Exhibit 1022, Page 60 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

21.4 Two-Dimensional Modulations

Phase-Shift Keying (PSK)

This is a linear modulation scheme generating the signal

$$v(t) = \Re\left\{ \sum_{n=-\infty}^{\infty} \xi_n s(t - nT) e^{j2\pi f_0 t} \right\}$$

where $\xi_n = e^{j\Phi_n}$ takes values in the set

$$\left\{ \frac{2\pi}{M}(i-1) + \Phi \right\}_{i=1}^{M}$$

where Φ is an arbitrary phase. The signal constellation is shown, for $M = 8$, in Fig. 21.4.

The symbol-error probability of M-ary PSK is closely approximated by

$$P(e) \approx \operatorname{erfc}\left(\sqrt{\frac{\mathcal{E}_b}{N_0}} \log_2 M \sin \frac{\pi}{M} \right)$$

for high signal-to-noise ratios: (See Fig. 21.5). The bandwidth efficiency of PSK is

$$\frac{R_s}{B} = \frac{\log_2 M}{T} \cdot T = \log_2 M$$

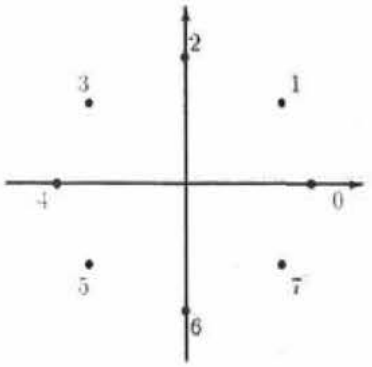


FIGURE 21.4 An 8-PSK constellation.

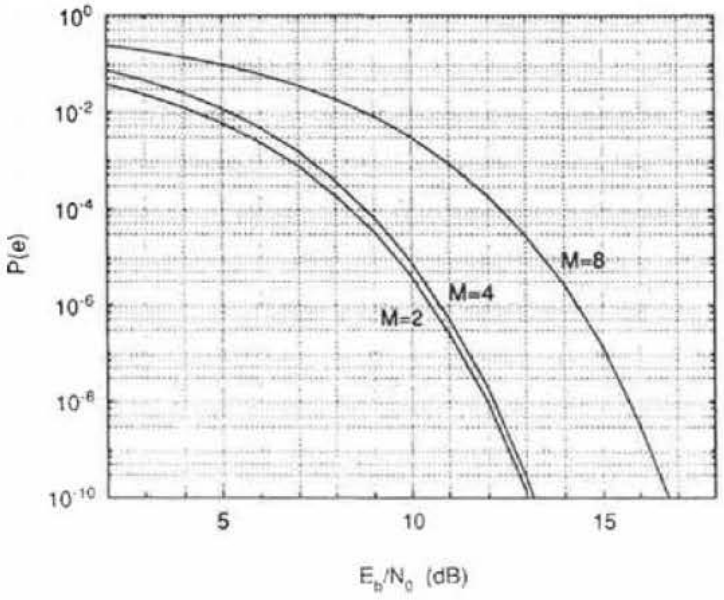


FIGURE 21.5 Symbol error probability of coherently demodulated PSK.

Samsung Exhibit 1022, Page 61 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Quadrature Amplitude Modulation (QAM)

Quadrature amplitude modulation (QAM) is a linear modulation scheme for which the modulated signal takes the form

$$v(t) = \Re\left\{ \sum_{n=-\infty}^{\infty} \xi_n s(t - nT) e^{j2\pi f_0 t} \right\}$$

where

$$\xi_n = \alpha_n + j\beta_n$$

and α_n, β_n take on equally spaced values. Figure 21.6 shows a QAM constellation with 16 points.

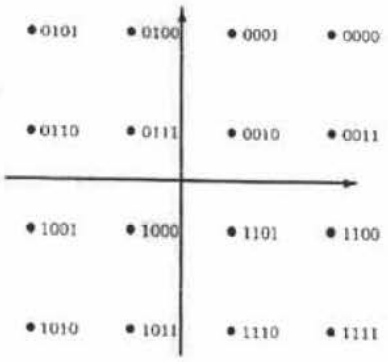


FIGURE 21.6 A 16-QAM constellation.

When $\log_2 M$ is an even integer, we have

$$P(e) = 1 - (1 - p)^2$$

with

$$p = \left(1 - \frac{1}{\sqrt{M}}\right) \operatorname{erfc}\left(\sqrt{\frac{3 \log_2 M}{2(M-1)} \frac{\mathcal{E}_b}{N_0}}\right)$$

When $\log_2 M$ is odd, the following upper bound holds:

$$P(e) < 2 \operatorname{erfc}\left(\sqrt{\frac{3 \log_2 M}{2(M-1)} \frac{\mathcal{E}_b}{N_0}}\right)$$

(See Fig. 21.7.)

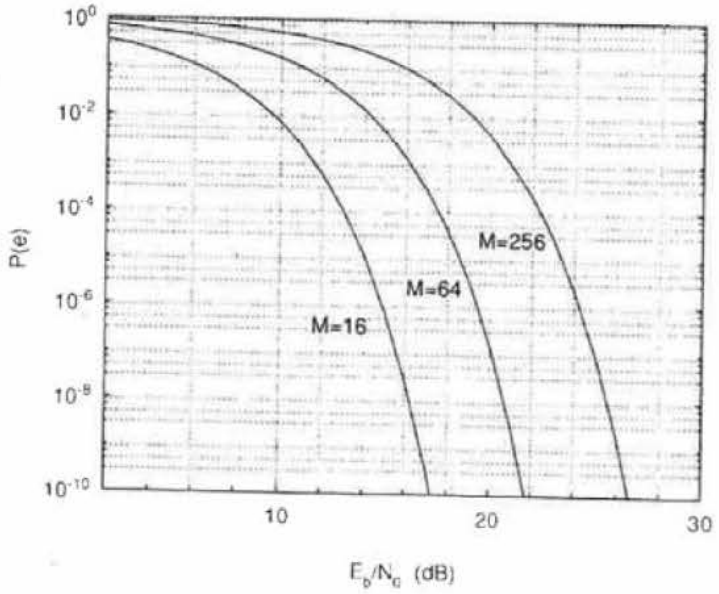


FIGURE 21.7 Symbol error probability of coherently demodulated QAM.

Samsung Exhibit 1022, Page 62 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

21.5 Multidimensional Modulations: Frequency-Shift Keying (FSK)

In the modulation technique called **frequency-shift keying** (FSK) the transmitter uses the waveforms

$$s_i(t) = A\cos 2\pi f_i t, \qquad 0 \le t < T$$

where $f_i = f_0 + (2i - 1 - M)f_d$, for $i = 1, \ldots, M$, and f_0 denotes the carrier frequency. The waveforms have constant energy $\mathcal{E} = A^2T/2$, and the modulated signal has constant envelope A. The distance properties of these waveforms for a given energy depend on the value of their normalized frequency separation $2f_dT$.

Generation of these waveforms may be accomplished with a set of M separate oscillators, each tuned to the frequency f_i. In this case, the phase discontinuities at frequency-transition times cause large sidelobes in the power density spectrum of the modulated signal. A narrower spectrum can be obtained by using a single oscillator whose frequency is modulated by the source bits. This results in a modulation scheme with memory called continuous-phase FSK (discussed in Sec. 21.7).

The signals $s_i(t)$ are orthogonal, that is,

$$\int_0^T \cos(2\pi f_i t)\cos(2\pi f_j t)\,dt = 0 \qquad i \ne j$$

if the frequency separation between adjacent signals is such that $2f_dT = m/2$, where m is any nonzero integer. Thus, the minimum frequency separation for orthogonality if $2f_d = 1/2T$.

The symbol error probability for M orthogonal signals with energy $\mathcal{E}$ is upper and lower bounded by

$$\frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{\mathcal{E}_b \log_2 M}{2N_0}}\right) \le P(e) \le \frac{M-1}{2}\operatorname{erfc}\left(\sqrt{\frac{\mathcal{E}_b \log_2 M}{2N_0}}\right).$$

For $M = 2$ the bounds merge into the exact value. The upper bound to $P(e)$ is plotted in Fig. 21.8.

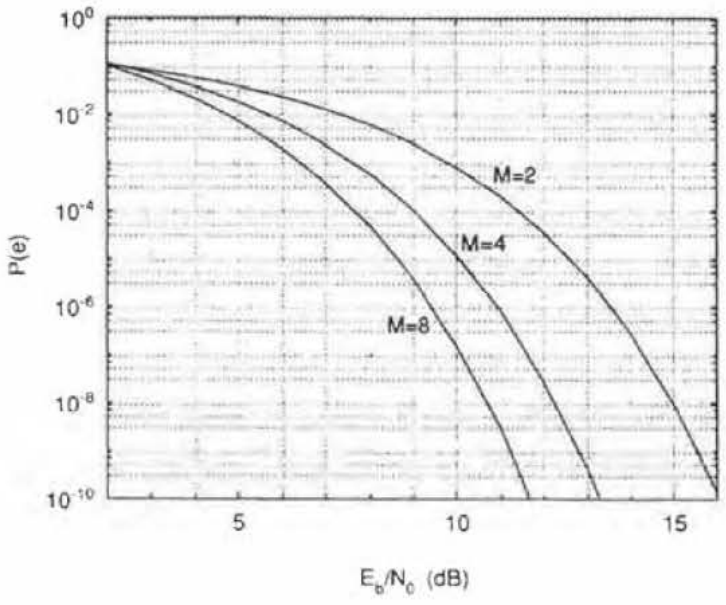


FIGURE 21.8 Symbol error probability of coherently demodulated orthogonal FSK.

Samsung Exhibit 1022, Page 63 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

It can be observed from Fig. 21.8 that the error probability for a given value of signal-to-noise ratio decreases as M increases, contrary to what is observed for PAM, PSK, and QAM. On the other hand, the bandwidth efficiency *decreases* as M increases, its value being given by

$$\frac{R_s}{B} = \frac{2\log_2 M}{M}$$

21.6 Multidimensional Modulations: Lattices

The QAM constellations may be considered as carved from an infinite constellation of regularly spaced points in the plane, called the **lattice** Z^2. This is represented in Fig. 21.9. This concept can be generalized to other lattices, both two dimensional and multi dimensional. In this framework, the design of a signal constellation consists of choosing a lattice, then carving a finite constellation of signal points from the lattice.

In general, a lattice Λ is defined as a set closed under ordinary addition and multiplication by integers. As such, it forms an additive group as well as a vector space. Specifically, an N-dimensional lattice Λ is an infinite set of N vectors.

If $d_{\min}$ is the minimum distance between any two points in the lattice, the number ν of adjacent lattice points located at distance $d_{\min}$, that is, the number of nearest neighbors of any lattice point, is called its *kissing number*.

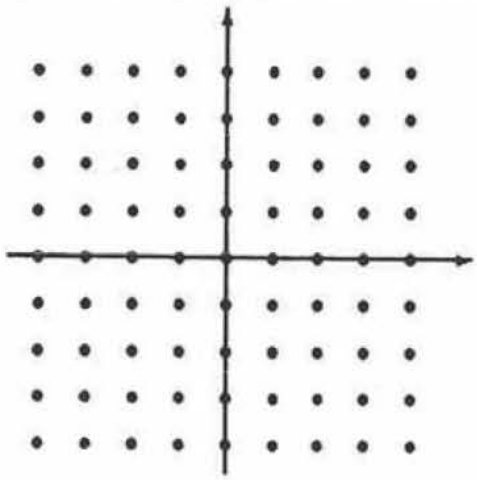


FIGURE 21.9 The lattice Z^2.

Examples of Lattices

The Lattice Z^N

The set Z^N of all N-tuples with integer coordinates is called the *cubic lattice*, or *integer lattice*. The minimum distance is $d_{\min} = 1$, and the kissing number is $\nu = 2N$.

The N-Dimensional Lattices A_N and A_N^*

A_N is the set of all vectors with $(N+1)$ integer coordinates whose sum is zero. This lattice may be viewed as the intersection of Z^{N+1} and a hyperplane cutting the origin. The minimum distance is $d_{\min} = \sqrt{2}$, and the kissing number is $\nu = N(N+1)$.

The N-Dimensional Lattice D_N

D_N is the set of all N-dimensional points whose integer coordinates have an even sum. It may be viewed as a punctured version of Z^N, in which the points are colored alternately black and white with a checkerboard coloring, and the white points (those with odd sums) are removed. We have $d_{\min} = \sqrt{2}$, and $\nu = 2N(N-1)$. D_4 represents the densest lattice packing in R^4. This means that if unit-radius, four-dimensional spheres with centers in the lattice points are used to pack R^4, then D_4 is the lattice with the largest number of spheres per unit volume.

Samsung Exhibit 1022, Page 64 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

smoother and the spectrum occupancy of the modulated signal is reduced. This modulation scheme is employed in the Global System for Mobile Communications (GSM) standard for cellular mobile radio with the choice $BT = 0.3$.

Defining Terms

Bandwidth: The frequency interval in which the density spectrum of a signal is significantly different from zero.

Continuous-phase modulation: A digital modulation scheme derived from phase-shift keying in which the carrier has no phase jumps, that is, the phase transitions occur continuously.

Digital modulation: The mapping of information-source symbols into signals, performed to carry information through the transmission channel.

Error probability: Probability that a symbol emitted by the information source will be received incorrectly by the end user.

Frequency-shift keying: A digital modulation scheme in which the source information is carried by the frequency of a sinusoidal waveform.

Geometric signal representation: Representation of a finite set of signals as a set of vector.

Lattice: An infinite signal constellation of points regularly located in space.

Phase-shift keying: A digital modulation scheme in which the source information is carried by the phase of a sinusoidal waveform, called the carrier.

Pulse-amplitude modulation: A digital modulation scheme in which the source information is carried by the amplitude of a waveform.

Quadrature amplitude modulation: A digital modulation scheme in which the source information is carried by the amplitude and by the phase of a sinusoidal waveform.

Signal constellation: A set of signals geometrically represented in the form of a set of vectors.

Signal-to-noise ratio: The ratio of the signal power and noise power. It is an index of channel quality.

References

Anderson, J.B., Aulin, T., and Sundberg, C.-E.W. 1986. *Digital Phase Modulation*, Plenum, New York.

Benedetto, S., Biglieri, E., and Castellani, V. 1987. *Digital Transmission Theory*, Prentice–Hall, Englewood Cliffs, NJ.

Conway, J.H. and Sloane, N.J.A. 1988. *Sphere Packings, Lattices and Groups*, Springer–Verlag, New York.

Forney, G.D. Jr. and Wei, L.-F. 1990. Multidimensional constellations—Part I: Introduction, figures of merit, and generalized cross constellations. *IEEE J. Selected Areas Comm.*, 7(6):877–892.

Massey, J.L. 1995. Towards an information theory of spread-spectrum systems. In: *Code Division Multiple Access Communications*, eds. S.G. Glisic and P.A. Leppänen, Kluwer Academic, Boston.

Proakis, J.G. 1995. *Digital Communications*, 3rd ed., McGraw–Hill, New York.

Simon, M.K., Hinedi, S.M., and Lindsey, W.C. 1995. *Digital Communication Techniques. Signal Design and Detection*, Prentice–Hall, Englewood Cliffs, NJ.

Slepian, D. 1976. On bandwidth. *IEEE Proc.*, 64(2):292–300.

Further Information

The monthly journal *IEEE Transactions on Communications* reports advances in digital modulation techniques.

The books by Benedetto, Biglieri, and Castellani 1987, Proakis 1995, and Simon, Hinedi, and Lindsey 1995 are good introductions to the theory of digital modulation.

Samsung Exhibit 1022, Page 65 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Horizontal Eye Opening

The width of each opening indicates the sensitivity to timing offset. Specifically, a very narrow eye opening indicates that a small timing offset will result in sampling where the eye is closed. Conversely, a wide horizontal opening indicates that a large timing offset can be tolerated, although the error probability will depend on the vertical opening.

Slope of the Inner Eye

The slope of the inner eye indicates sensitivity to timing jitter or variance in the timing offset. Specifically, a very steep slope means that the eye closes rapidly as the timing offset increases. In this case, a significant amount of jitter in the sampling times significantly increases the probability of error.

The shape of the eye diagram is determined by the pulse shape. In general, the faster the baseband pulse decays, the wider the eye opening. For example, a rectangular pulse produces a box-shaped eye diagram (assuming binary signalling). The minimum bandwidth pulse shape Eq. (25.12) produces an eye diagram which is closed for all t except for $t = 0$. This is because, as shown earlier, an arbitrarily small timing offset can lead to an intersymbol interference term that is arbitrarily large, depending on the data sequence.

25.5 Partial-Response Signalling

To avoid the problems associated with Nyquist signalling over an ideal bandlimited channel, bandwidth and/or power efficiency must be compromised. Raised cosine pulses compromise bandwidth efficiency to gain robustness with respect to timing errors. Another possibility is to introduce a controlled amount of intersymbol at the transmitter, which can be removed at the receiver. This approach is called **partial-response (PR) signalling**. The terminology reflects the fact that the sampled system impulse response does not have the full response given by the Nyquist condition Eq. (25.7).

To illustrate PR signalling, suppose that the Nyquist condition Eq. (25.7) is replaced by the condition

$$h_k = \begin{cases} 1 & k = 0, 1 \\ 0 & \text{all other } k \end{cases} \tag{25.26}$$

The kth received sample is then

$$y_k = A_k + A_{k-1} + \tilde{n}_k \tag{25.27}$$

so that there is intersymbol interference from one neighboring transmitted symbol. For now we focus on the spectral characteristics of PR signalling and defer discussion of how to detect the transmitted sequence $\{A_k\}$ in the presence of intersymbol interference. The equivalent discrete-time transfer function in this case is the discrete Fourier transform of the sequence in Eq. (25.26),

$$\begin{aligned} H_{\text{eq}}(e^{j2\pi fT}) &= \frac{1}{T}\sum_k H\left(f + \frac{k}{T}\right) \\ &= 1 + e^{-j2\pi fT} = 2e^{-j\pi fT}\cos(\pi fT) \end{aligned} \tag{25.28}$$

As in the full-response case, for Eq. (25.28) to be satisfied, the *minimum* bandwidth of the channel $G(f)$ and transmitter filter $P(f)$ is $W = 1/(2T)$. Assuming $P(f)$ has this minimum bandwidth

Samsung Exhibit 1022, Page 66 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

implies

$$H(f) = \begin{cases} 2Te^{-j\pi fT}\cos(\pi fT) & |f| < 1/(2T) \\ 0 & |f| > 1/(2T) \end{cases} \tag{25.29a}$$

and

$$h(t) = T\{\text{sinc}(t/T) + \text{sinc}[(t-T)/T]\} \tag{25.29b}$$

where $\text{sinc}\, x = (\sin \pi x)/(\pi x)$. This pulse is called a *duobinary* pulse and is shown along with the associated $H(f)$ in Fig. 25.9. [Notice that $h(t)$ satisfies Eq. (25.26).] Unlike the ideal bandlimited frequency response, the transfer function $H(f)$ in Eq. (25.29a) is continuous and is, therefore, easily approximated by a physically realizable filter. Duobinary PR was first proposed by Lender, 1963, and later generalized by Kretzmer, 1966.

The main advantage of the duobinary pulse Eq. (25.29b), relative to the minimum bandwidth pulse Eq. (25.12), is that signalling at the Nyquist symbol rate is feasible with zero excess bandwidth. Because the pulse decays much more rapidly than a Nyquist pulse, it is robust with respect to timing errors. Selecting the transmitter and receiver filters so that the overall system response is duobinary is appropriate in situations where the channel frequency response $G(f)$ is near zero or has a rapid rolloff at the Nyquist band edge $f = 1/(2T)$.

As another example of PR signaling, consider the *modified* duobinary partial response

$$h_k = \begin{cases} 1 & k = -1 \\ -1 & k = 1 \\ 0 & \text{all other } k \end{cases} \tag{25.30}$$

which has equivalent discrete-time transfer function

$$\begin{aligned} H_{\text{eq}}(e^{j2\pi fT}) &= e^{j2\pi fT} - e^{-j2\pi fT} \\ &= j2\sin(2\pi fT) \end{aligned} \tag{25.31}$$

With zero excess bandwidth, the overall system response is

$$H(f) = \begin{cases} j2T\sin(2\pi fT) & |f| < 1/(2T) \\ 0 & |f| > 1/(2T) \end{cases} \tag{25.32a}$$

and

$$h(t) = T\{\text{sinc}[(t+T)/T] - \text{sinc}[(t-T)/T]\} \tag{25.32b}$$

These functions are plotted in Fig. 25.10. This pulse shape is appropriate when the channel response $G(f)$ is near zero at both DC ($f = 0$) and at the Nyquist band edge. This is often the case for wire (twisted-pair) channels where the transmitted signal is coupled to the channel through a transformer. Like duobinary PR, modified duobinary allows minimum bandwidth signalling at the Nyquist rate.

A particular partial response is often identified by the polynomial

$$\sum_{k=0}^{K} h_k D^k$$

where D (for delay) takes the place of the usual z^{-1} in the z transform of the sequence $\{h_k\}$. For example, duobinary is also referred to as $1 + D$ partial response.

Samsung Exhibit 1022, Page 67 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

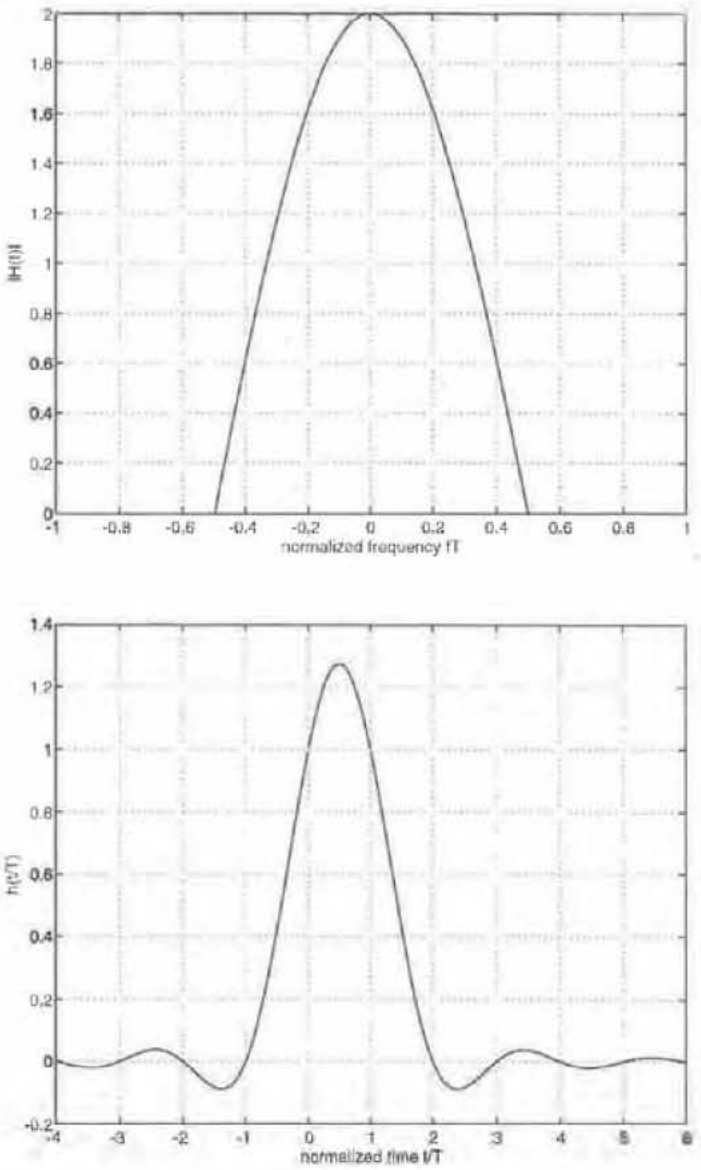


FIGURE 25.9 Duobinary frequency response and minimum bandwidth pulse.

In general, more complicated system responses than those shown in Figs. 25.9 and 25.10 can be generated by choosing more nonzero coefficients in the sequence $\{h_k\}$. This complicates detection, however, because of the additional intersymbol interference that is generated.

Rather than modulating a PR pulse $h(t)$, a PR signal can also be generated by filtering the sequence of transmitted levels $\{A_i\}$. This is shown in Fig. 25.11. Namely, the transmitted levels are first passed through a discrete-time (digital) filter with transfer function $P_d(e^{j2\pi fT})$ (where the subscript d indicates discrete). [Note that $P_d(e^{j2\pi fT})$ can be selected to be $H_{\text{eq}}(e^{j2\pi fT})$.] The outputs of this filter form the PAM signal, where the pulse shaping filter $P(f) = 1$, $|f| < 1/(2T)$ and is zero elsewhere. If the transmitted levels $\{A_k\}$ are selected independently and are identically distributed, then the transmitted spectrum is $\sigma_A^2 |P_d(e^{j2\pi fT})|^2$ for $|f| < 1/(2T)$ and is zero for $|f| > 1/(2T)$, where $\sigma_A^2 = E[|A_k|^2]$.

Samsung Exhibit 1022, Page 68 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

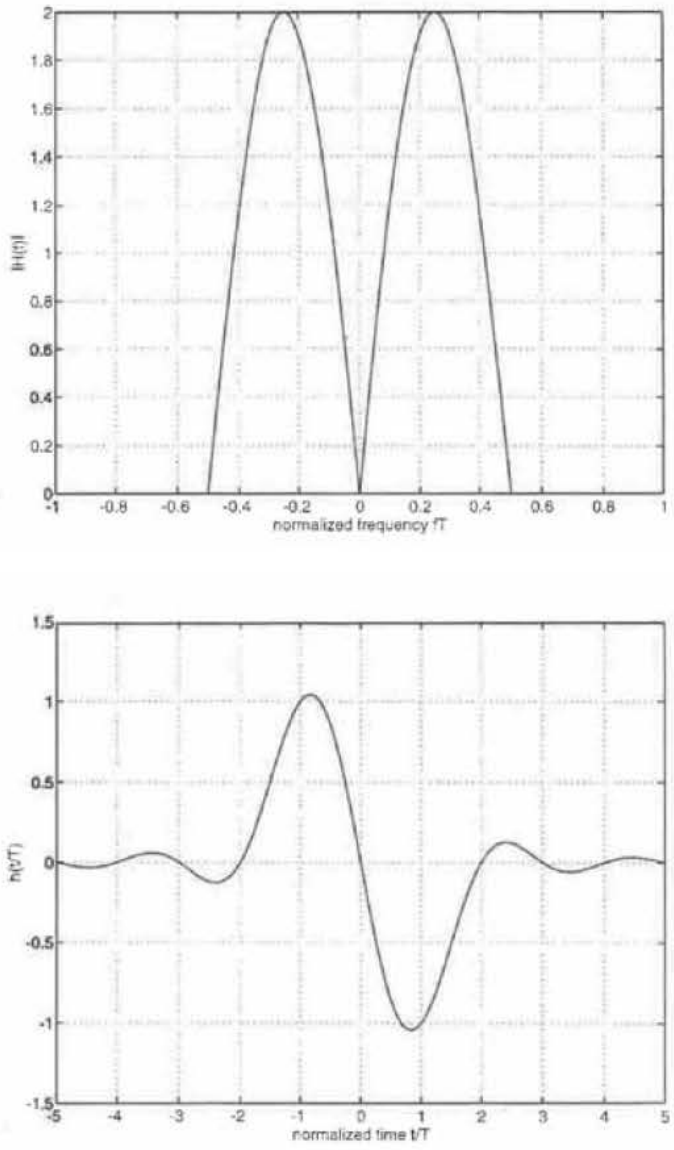


FIGURE 25.10 Modified duobinary frequency response and minimum bandwidth pulse.

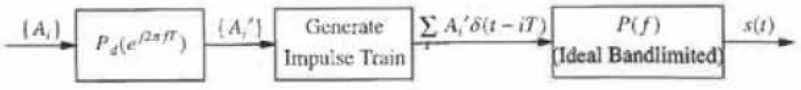


FIGURE 25.11 Generation of PR signal.

Shaping the transmitted spectrum to have nulls coincident with nulls in the channel response potentially offers significant performance advantages. By introducing intersymbol interference, however, PR signalling increases the number of received signal levels, which increases the complexity of the detector and may reduce immunity to noise. For example, the set of received signal levels for duobinary signalling is $\{0, \pm 2\}$ from which the transmitted levels $\{\pm 1\}$ must be estimated. The

Samsung Exhibit 1022, Page 69 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540



FIGURE 25.12 Precoding for a PR channel.

TABLE 25.1 Example of Precoding for Duobinary PR.

$\{b_i\}$:		1	0	0	1	1	1	0	0	1	0
$\{b_i'\}$:	0	1	1	1	0	1	0	0	0	1	1
$\{A_i\}$:	−1	1	1	1	−1	1	−1	−1	−1	1	1
$\{y_i\}$:		0	2	2	0	0	0	−2	−2	0	2

performance of a particular PR scheme depends on the channel characteristics, as well as the type of detector used at the receiver. We now describe a simple suboptimal detection strategy.

Precoding

Consider the received signal sample Eq. (25.27) with duobinary signalling. If the receiver has correctly decoded the symbol A_{k-1}, then in the absence of noise A_k can be decoded by subtracting A_{k-1} from the received sample y_k. If an error occurs, however, then subtracting the preceding symbol estimate from the received sample will cause the error to propogate to successive detected symbols. To avoid this problem, the transmitted levels can be **precoded** in such a way as to compensate for the intersymbol interference introduced by the overall partial response.

We first illustrate precoding for duobinary PR. The sequence of operations is illustrated in Fig. 25.12. Let $\{b_k\}$ denote the sequence of source bits where $b_k \in \{0, 1\}$. This sequence is transformed to the sequence $\{b_k'\}$ by the operation

$$b_k' = b_k \oplus b_{k-1}' \tag{25.33}$$

where $\oplus$ denotes modulo 2 addition (exclusive OR). The sequence $\{b_k'\}$ is mapped to the sequence of binary transmitted signal levels $\{A_k\}$ according to

$$A_k = 2b_k' - 1 \tag{25.34}$$

That is, $b_k' = 0$ ($b_k = 1$) is mapped to the transmitted level $A_k = -1$ ($A_k = 1$). In the absence of noise, the received symbol is then

$$y_k = A_k + A_{k-1} = 2(b_k' + b_{k-1}' - 1) \tag{25.35}$$

and combining Eqs. (25.33) and (25.35) gives

$$b_k = \left(\frac{1}{2}y_k + 1\right) \bmod 2 \tag{25.36}$$

That is, if $y_k = \pm 2$, then $b_k = 0$, and if $y_k = 0$, then $b_k = 1$. Precoding, therefore, enables the detector to make *symbol-by-symbol* decisions that do not depend on previous decisions. Table 25.1 shows a sequence of transmitted bits $\{b_i\}$, precoded bits $\{b_i'\}$, transmitted signal levels $\{A_i\}$, and received samples $\{y_i\}$.

Samsung Exhibit 1022, Page 70 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The preceding precoding technique can be extended to multilevel PAM and to other PR channels. Suppose that the PR is specified by

$$H_{\text{eq}}(D) = \sum_{k=0}^{K} h_k D^k$$

where the coefficients are integers and that the source symbols $\{b_k\}$ are selected from the set $\{0, 1, \ldots, M-1\}$. These symbols are transformed to the sequence $\{b'_k\}$ via the precoding operation

$$b'_k = \left(b_k - \sum_{i=1}^{K} h_i b'_{k-i} \right) \bmod M \tag{25.37}$$

Because of the modulo operation, each symbol b'_k is also in the set $\{0, 1, \ldots, M-1\}$. The kth transmitted signal level is given by

$$A_k = 2b'_k - (M-1) \tag{25.38}$$

so that the set of transmitted levels is $\{-(M-1), \ldots, (M-1)\}$ (i.e., a shifted version of the set of values assumed by b_k). In the absence of noise the received sample is

$$y_k = \sum_{i=0}^{K} h_i A_{k-i} \tag{25.39}$$

and it can be shown that the kth source symbol is given by

$$b_k = \frac{1}{2}(y_k + (M-1) \cdot H_{\text{eq}}(1)) \bmod M \tag{25.40}$$

Precoding the symbols $\{b_k\}$ in this manner, therefore, enables symbol-by-symbol decisions at the receiver. In the presence of noise, more sophisticated detection schemes (e.g., maximum likelihood) can be used with PR signalling to obtain improvements in performance.

25.6 Additional Considerations

In many applications, bandwidth and intersymbol interference are not the only important considerations for selecting baseband pulses. Here we give a brief discussion of additional practical constraints that may influence this selection.

Average Transmitted Power and Spectral Constraints

The constraint on average transmitted power varies according to the application. For example, low-average power is highly desirable for mobile wireless applications that use battery-powered transmitters. In many applications (e.g., digital subscriber loops, as well as digital radio), constraints are imposed to limit the amount of interference, or crosstalk, radiated into neighboring receivers and communications systems. Because this type of interference is frequency dependent, the constraint may take the form of a spectral mask that specifies the maximum allowable transmitted power as a function of frequency. For example, crosstalk in wireline channels is generally caused by capacitive coupling and increases as a function of frequency. Consequently, to reduce the amount of crosstalk generated at a particular transmitter, the pulse shaping filter generally attenuates high frequencies more than low frequencies.

Samsung Exhibit 1022, Page 71 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

In radio applications where signals are assigned different frequency bands, constraints on the transmitted spectrum are imposed to limit *adjacent-channel interference.* This interference is generated by transmitters assigned to adjacent frequency bands. Therefore, a constraint is needed to limit the amount of *out-of-band power* generated by each transmitter, in addition to an overall average power constraint. To meet this constraint, the transmitter filter in Fig. 25.3 must have a sufficiently steep rolloff at the edges of the assigned frequency band. (Conversely, if the transmitted signals are time multiplexed, then the duration of the system impulse response must be contained within the assigned time slot.)

Peak-to-Average Power

In addition to a constraint on average transmitted power, a *peak-power* constraint is often imposed as well. This constraint is important in practice for the following reasons:

1. The dynamic range of the transmitter is limited. In particular, saturation of the output amplifier will "clip" the transmitted waveform.
2. Rapid fades can severely distort signals with high peak-to-average power.
3. The transmitted signal may be subjected to nonlinearities. Saturation of the output amplifier is one example. Another example that pertains to wireline applications is the companding process in the voice telephone network [Kalet and Saltzberg, 1994]. Namely, the compander used to reduce quantization noise for pulse-code modulated voice signals introduces amplitude-dependent distortion in data signals.

The preceding impairments or constraints indicate that the transmitted waveform should have a low peak-to-average power ratio. The peak-to-average power ratio is minimized by using binary signalling with rectangular pulse shapes. However, this compromises bandwidth efficiency. In applications where peak-to-average ratio should be low, binary signalling with rounded pulses are often used.

Channel and Receiver Characteristics

The type of channel impairments encountered and the type of detection scheme used at the receiver can also influence the choice of a transmitted pulse shape. For example, a constant amplitude pulse is appropriate for a fast fading environment with noncoherent detection. The ability to track channel characteristics, such as phase, may allow more bandwidth efficient pulse shapes in addition to multilevel signalling.

High-speed data communications over time-varying channels requires that the transmitter and/or receiver adapt to the changing channel characteristics. Adapting the transmitter to compensate for a time-varying channel requires a feedback channel through which the receiver can notify the transmitter of changes in channel characteristics. Because of this extra complication, adapting the receiver is often preferred to adapting the transmitter pulse shape. However, the following examples are notable exceptions.

1. The current IS-95 air interface for direct-sequence code-division multiple access adapts the transmitter power to control the amount of interference generated and to compensate for channel fades. This can be viewed as a simple form of adaptive transmitter pulse shaping in which a single parameter associated with the pulse shape is varied.
2. Multitone modulation divides the channel bandwidth into small subbands, and the transmitted power and source bits are distributed among these subbands to maximize the information rate. The received signal-to-noise ratio for each subband must be transmitted back to the transmitter to guide the allocation of transmitted bits and power [Bingham, 1990].

In addition to multitone modulation, *adaptive precoding* (also known as Tomlinson–Harashima precoding [Tomlinson, 1971; Harashima and Miyakawa, 1972]) is another way in which the trans-

Samsung Exhibit 1022, Page 72 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

mitter can adapt to the channel frequency response. Adaptive precoding is an extension of the technique described earlier for partial-response channels. Namely, the equivalent discrete-time channel impulse response is measured at the receiver and sent back to the transmitter, where it is used in a precoder. The precoder compensates for the intersymbol interference introduced by the channel, allowing the receiver to detect the data by a simple threshhold operation. Both multitone modulation and precoding have been used with wireline channels (voiceband modems and digital subscriber loops).

Complexity

Generation of a bandwidth-efficient signal requires a filter with a sharp cutoff. In addition, bandwidth-efficient pulse shapes can complicate other system functions, such as timing and carrier recovery. If sufficient bandwidth is available, the cost can be reduced by using a rectangular pulse shape with a simple detection strategy (low-pass filter and threshold).

Tolerance to Interference

Interference is one of the primary channel impairments associated with digital radio. In addition to adjacent-channel interference described earlier, *cochannel interference* may be generated by other transmitters assigned to the same frequency band as the desired signal. Co-channel interference can be controlled through frequency (and perhaps time slot) assignments and by pulse shaping. For example, assuming fixed average power, increasing the bandwidth occupied by the signal lowers the power spectral density and decreases the amount of interference into a narrowband system that occupies part of the available bandwidth. Sufficient bandwidth spreading, therefore, enables wideband signals to be overlaid on top of narrowband signals without disrupting either service.

Probability of Intercept and Detection

The broadcast nature of wireless channels generally makes eavesdropping easier than for wired channels. A requirement for most commercial, as well as military applications, is to guarantee the privacy of user conversations (low probability of intercept). An additional requirement, in some applications, is that determining whether or not communications is taking place must be difficult (low probability of detection). Spread-spectrum waveforms are attractive in these applications since spreading the pulse energy over a wide frequency band decreases the power spectral density and, hence, makes the signal less visible. Power-efficient modulation combined with coding enables a further reduction in transmitted power for a target error rate.

25.7 Examples

We conclude this chapter with a brief description of baseband pulse shapes used in existing and emerging standards for digital mobile cellular and Personal Communications Services (PCS).

Global System for Mobile Communications (GSM)

The European GSM standard for digital mobile cellular communications operates in the 900-MHz frequency band, and is based on time-division multiple access (TDMA) [Rahnema, 1993]. A special variant of binary FSK is used called *Gaussian minimum-shift keying (GMSK)*. The GMSK modulator is illustrated in Fig. 25.13. The input to the modulator is a binary PAM signal $s(t)$, given by Eq. (25.3), where the pulse $p(t)$ is a Gaussian function and $|s(t)| < 1$. This waveform frequency modulates the carrier f_c, so that the (passband) transmitted signal is

$$w(t) = K\cos\left[2\pi f_c t + 2\pi f_d \int_{-\infty}^{t} s(\tau)\,\mathrm{d}\tau\right]$$

Samsung Exhibit 1022, Page 73 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540



FIGURE 25.13 Generation of GMSK signal; LPF is low-pass filter.

The maximum frequency deviation from the carrier is $f_d = 1/(2T)$, which characterizes minimum-shift keying. This technique can be used with a noncoherent receiver that is easy to implement. Because the transmitted signal has a constant envelope, the data can be reliably detected in the presence of rapid fades that are characteristic of mobile radio channels.

U.S. Digital Cellular (IS-54)

The IS-54 air interface operates in the 800-MHz band and is based on TDMA [EIA/TIA, 1991]. The baseband signal is given by Eq. (25.3) where the symbols are complex-valued, corresponding to quadrature phase modulation. The pulse has a square-root raised cosine spectrum with 35% excess bandwidth.

Interim Standard-95

The IS-95 air interface for digital mobile cellular uses spread-spectrum signalling (CDMA) in the 800-MHz band [TIA, 1993]. The baseband transmitted pulse shapes are analogous to those shown in Fig. 25.2, where the number of square pulses (chips) per bit is 128. To improve spectral efficiency the (wideband) transmitted signal is filtered by an approximation to an ideal low-pass response with a small amount of excess bandwidth. This shapes the chips so that they resemble minimum bandwidth pulses.

Personal Access Communications System (PACS)

Both PACS and the Japanese personal handy phone (PHP) system are TDMA systems which have been proposed for personal communications systems (PCS), and operate near 2 Ghz [Cox, 1995]. The baseband signal is given by Eq. (25.3) with four complex symbols representing four-phase quadrature modulation. The baseband pulse has a square-root raised cosine spectrum with 50% excess bandwidth.

Defining Terms

Baseband signal: A signal with frequency content centered around DC.

Equivalent discrete-time transfer function: A discrete-time transfer function (z transform) that relates the transmitted amplitudes to received samples in the absence of noise.

Excess bandwidth: That part of the baseband transmitted spectrum which is not contained within the Nyquist band.

Eye diagram: Superposition of segments of a received PAM signal that indicates the amount of intersymbol interference present.

Frequency-shift keying: A digital modulation technique in which the transmitted pulse is sinusoidal, where the frequency is determined by the source bits.

Intersymbol interference: The additive contribution (interference) to a received sample from transmitted symbols other than the symbol to be detected.

Samsung Exhibit 1022, Page 74 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Matched filter: The receiver filter with impulse response equal to the time-reversed, complex conjugate impulse response of the combined transmitter filter-channel impulse response.

Nyquist band: The narrowest frequency band that can support a PAM signal without intersymbol interference (the interval $[-1/(2T), 1/(2T)]$ where $1/T$ is the symbol rate).

Nyquist criterion: A condition on the overall frequency response of a PAM system that ensures the absence of intersymbol interference.

Partial-response signalling: A signalling technique in which a controlled amount of intersymbol interference is introduced at the transmitter in order to shape the transmitted spectrum.

Precoding: A transformation of source symbols at the transmitter that compensates for intersymbol interference introduced by the channel.

Pulse amplitude modulation (PAM): A digital modulation technique in which the source bits are mapped to a sequence of amplitudes that modulate a transmitted pulse.

Raised cosine pulse: A pulse shape with Fourier transform that decays to zero according to a raised cosine; see Eq. (25.18). The amount of excess bandwidth is conveniently determined by a single parameter (α).

Spread spectrum: A signalling technique in which the pulse bandwidth is many times wider than the Nyquist bandwidth.

Zero-forcing criterion: A design constraint which specifies that intersymbol interference be eliminated.

References

Bingham, J.A.C. 1990. Multicarrier modulation for data transmission: an idea whose time has come. *IEEE Commun. Mag.* 28(May):5–14.

Cox, D.C. 1995. Wireless personal communications: what is it? *IEEE Personal Comm.* 2(2):20–35.

Electronic Industries Association/Telecommunications Industry Association. 1991. Recommended minimum performance standards for 800 MHz dual-mode mobile stations. Incorp. EIA/TIA 19B, EIA/TIA Project No. 2216, March.

Harashima, H. and Miyakawa, H. 1972. Matched-transmission technique for channels with intersymbol interference. *IEEE Trans. on Commun.* COM-20(Aug.):774–780.

Kalet, I. and Saltzberg, B.R. 1994. QAM transmission through a companding channel—signal constellations and detection. *IEEE Trans. on Comm.* 42(2–4):417–429.

Kretzmer, E.R. 1966. Generalization of a technique for binary data communication. *IEEE Trans. Comm. Tech.* COM-14 (Feb.):67, 68.

Lender, A. 1963. The duobinary technique for high-speed data Transmission. *AIEE Trans. on Comm. Electronics*, 82 (March):214–218.

Rahnema, M. 1993. Overview of the GSM system and protocol architecture. *IEEE Commun. Mag.* (April):92–100.

Telecommunication Industry Association. 1993. Mobile station-base station compatibility standard for dual-mode wideband spread spectrum cellular system. TIA/EIA/IS-95. July.

Tomlinson, M. 1971. New automatic equalizer employing modulo arithmetic. *Electron. Lett.* 7(March):138, 139.

Further Information

Baseband signalling and pulse shaping is fundamental to the design of any digital communications system and is, therefore, covered in numerous texts on digital communications. For more advanced treatments see E.A. Lee and D.G. Messerschmitt, *Digital Communication*, Kluwer 1994, and J.G. Proakis, *Digital Communications*, McGraw-Hill 1995.

Samsung Exhibit 1022, Page 75 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

32

Switching Fabrics

32.1 Introduction 425
32.2 Virtual Circuit Switching 426
32.3 Packet Switching 426
32.4 Crosspoints 426
32.5 Multiplexing Versus Divisions 427
Types of Divisions
32.6 Switching System Functions 428
32.7 Topology of Switch Fabrics 428
32.8 Concentration, Expansion and Distribution 428
32.9 Multistage Fabrics 429
32.10 Fabric Control 429
32.11 Topology 430
Time-Division Switching Elements • Time Multiplex Switching (TMS) • Time-Slot Interchange (TSI) • Time-Division Switch Fabrics
32.12 Fabrics for Asynchronous Transfer Mode (ATM) 432

Amos E. Joel Jr.

32.1 Introduction

There are two distinct principles employed in electronic switching systems. They are known as *circuit switching* and *packet switching*.

In circuit switching a dedicated path is selected, established, and utilized for the entire duration a message session between a calling terminal and at least one selected called terminal. It usually has the capability of simultaneous transmission in both directions, known as duplex. Establishment of a channel through one or more switches in sequence is known as a *connection*. The process of requesting and establishing a connection is known as a *call*. The bandwidth associated with each service of a call is reserved for the duration of the service and is dedicated to that call.

In telecommunication systems the most important element is the message. If there are only two terminals, one to originate and another to terminate messages, no switching is required. When there are more than two terminals, unless messages are broadcast, switching is required to provide *selective* communication among a number of terminals. Generally privacy is required so that a message will be received by only the terminal for which the message is intended. Communication privacy requires that *contention* be included in the selection process to insure privacy and that a message reaches only the selected terminal over the selected path. *Selection* and *contention* are basic to switching.

Contention for the selection of channels occurs only at the time when a connection is being established. If a complete path is not possible when requested, the calling terminal is usually requested or scheduled to try to establish the connection again later.

0-8493-8349-8/97/$0.00+$.50

© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 76 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The function within a switch that provides paths for connections is known as the *switch fabric.* This term is relatively new. In the past it has been known as the switching network or the switching center network.

In circuit switches, a message may be sent only one way, directed only from one terminal to another, or there may be interactive or two-way messages. For two-way messages that occur in real time, circuit switching systems establish a path between two or more terminals exchanging messages.[1]

32.2 Virtual Circuit Switching

Circuit switches are generally transparent to message content. However, to obtain efficiencies, circuit switching systems may remove or disconnect an established path in either or both directions during silent periods in the message and re-establish the circuit when the message information resumes. This technique is sometimes referred to as *virtual circuit switching.* Virtual circuit switching assumes that there is no perceptible degradation of the message. This perception depends on the service quality expectations of users.

The combination of transmission, switching, and terminals form a telecommunications *network.* Where the nature of messages may not require real-time interaction, other forms of networks with different economic criteria for service quality and message delay may be employed. They may also be used where the network serves more than one type of telecommunications, for example, voice, data, and video.

32.3 Packet Switching

In packet switching, messages are divided into segments or packets. These packets may be transmitted over different selected channels in a network. Each packet contains address information as well as other information regarding message treatment. Contention occurs for each packet at each switch. Most applications using packet switching are, traditionally at least, one way and not necessarily in real time; packets may be stored and delayed until the contention is resolved or the switch runs out of storage capacity. The memory used to store the packets is, for packet switching, the equivalent of the switching fabrics.[2]

32.4 Crosspoints

There are two basic technology elements used in switch fabrics. One is the memory previously referred to and the other is simply a switch often referred to as a *crosspoint.* Any device that has at least two stable states, such as one that opens and closes a circuit or changes from a high to low electrical impedance may be used as a crosspoint. Crosspoints are assembled in various ways to the form switch fabrics. The following are some technologies that have been used:

- Manual devices, for example, plugs and jacks. The plugs may be associated with flexible cords, several of which operate over a field of jacks that together constitute a fabric.
- Gross (large) motion electromechanical devices, for example, two motion (up and around) step-by-step switches and those that move in only one direction, such as rotary selectors and panel selectors that move over a vertical plane. In gross motion switches, wipers or brushes move over fixed banks of contacts. The combination of wiper and bank constitutes crosspoints.

[1]Circuit switching channels may also be established on a schedule.

[2]Asynchronous transfer mode (ATM) is a form of packet transmission with provision for real-time interactive (voice) messaging. The packets are of uniform length called *cells.* Switch fabrics for ATM (discussed subsequently) may employ crosspoints as well as memory.

Samsung Exhibit 1022, Page 77 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

41

The Open Systems Interconnections (OSI) Seven-Layer Model

41.1 Computer Communications Requirements 567
41.2 Standards Evolution 568
41.3 International Standards Organization Reference Model 569
The Application-Oriented Layers • The Presentation Layer • The Network-Dependent Layers
41.4 Open System Standards 574
41.5 Summary 576

Fred Halsall
University of Wales, Swansea

41.1 Computer Communications Requirements

Although in many instances computers are used to perform their intended role in a stand-alone mode, in others there is a need to interwork and exchange data with other computers. In financial applications, for example, to carry out funds transfers from one institution computer to another, in travel applications to access the reservation systems belonging to various airlines, and so on. The general requirement in all of these applications is for application programs running in different computers to cooperate to achieve a specific distributed application function. To achieve this, three basic issues must be considered. These are shown in diagrammatic form in Fig. 41.1.

The fundamental requirement in all applications that involve two or more computers is the provision of a suitable data communications facility. This may comprise a **local area network** (LAN), if the computers are distributed around a single site, a **wide area network** (WAN), if the computers are situated at different sites or an **internetwork** if multiple interconnected network types are involved. Associated with these different network types is a set of access protocols which enable a communications path between two computers to be established and for data to be transferred across this path. Typically, these protocols differ for the different network types. In addition to these access protocols, the communication subsystem in each computer must provide additional functionality. For example, if the communicating computers are of different types, possibly with different word sizes and character sets, then a means of ensuring the transferred data is interpreted in the same way in each computer must be incorporated. Also, the computers may use different file systems and hence functionality to enable application programs, normally referred to as **application processes (APs)**, to access these in a standardized way must

0-8493-8349-8/97/$0.00+$.50

© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 78 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

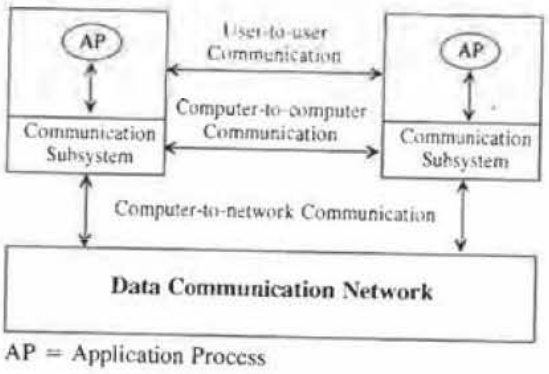


FIGURE 41.1 Computer communication schematic.

also be included. All of these issues must be considered when communicating data between two computers.

41.2 Standards Evolution

Until recently, the standards established for use in the computer industry by the various international bodies were concerned primarily with either the internal operation of a computer or the connection of a local peripheral device. The result was that early hardware and software communication subsystems offered by manufacturers only enables their own computers, and so-called *plug-compatible systems*, to exchange information. Such systems are known as *closed systems* since computers from other manufacturers cannot exchange information unless they adhere to the (proprietary) standards of a particular manufacturer.

In contrast, the various international bodies concerned with public-carrier telecommunication networks have for many years formulated internationally agreed standards for connecting devices to these networks. The *V-series recommendations*, for example, are concerned with the connection of equipment, normally referred to as a *data terminal equipment* (DTE), to a modem connected to the public switched telephone network (PSTN); the *X-series recommendations* for connecting a DTE to a public data network; and the *I-series recommendations* for connecting a DTE to the integrated services digital networks (ISDNs). The recommendations have resulted in compatibility between the equipment from different vendors, enabling a purchaser to select suitable equipment from a range of manufacturers.

Initially, the services provided by most public carriers were concerned primarily with data transmission, and hence the associated standards only related to the method of interfacing a device to these networks. More recently, however, the public carriers have started to provide more extensive distributed information services, such as the exchange of electronic messages (*teletex*) and access to public databases (*videotex*). To cater for such services, the standards bodies associated with the telecommunications industry have formulated standards not only for interfacing to such networks but also so-called higher level standards concerned with the format (syntax) and control of the exchange of information (data) between systems. Consequently, the equipment from one manufacturer that adheres to these standards can be interchangeable with equipment from any other manufacturer that complies with the standards. The resulting system is then known as an *open system* or, more completely, as an *open systems interconnection environment* (*OSIE*).

In the mid-1970s as different types of distributed systems (based on both public and private data networks) started to proliferate, the potential advantages of open systems were acknowledged by the computer industry. As a result, a range of standards started to be introduced. The first

Samsung Exhibit 1022, Page 79 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

was concerned with the overall structure of the complete communication subsystem within each computer. This was produced by the *International Standards Organization* (*ISO*) and is known as the *ISO reference model* for **open systems interconnection** (**OSI**).

The aim of the ISO reference model is to provide a framework for the coordination of standards development and to allow existing and evolving standards activities to be set within a common framework. The aim is to allow an application process in any computer that supports a particular set of standards to communication freely with an application process in any other computer that supports the same standards, irrespective of its origin of manufacture.

Some examples of application processes that may wish to communicate in an open way are the following:

- A process (program) executing in a computer and accessing a remote file system
- A process acting as a central file service (server) to a distributed community of (client) processes
- A process in an office workstation (computer) accessing an electronic mail service
- A process acting as an electronic mail server to a distributed community of (client) processes
- A process in a supervisory computer controlling a distributed community of computer-based instruments or robot controllers associated with a process or automated manufacturing plant
- A process in an instrument or robot controller receiving commands and returning results to a supervisory system
- A process in a bank computer that initiates debit and credit operations on a remote system

Open systems interconnection is concerned with the exchange of information between such processes. The aim is to enable application processes to cooperate in carrying out a particular (distributed) information processing task irrespective of the computers on which they are running.

41.3 International Standards Organization Reference Model

A communication subsystem is a complex piece of hardware and software. Early attempts at implementing the software for such subsystems were often based on a single, complex, unstructured program (normally written in assembly language) with many interacting components. The resulting software was difficult to test and often very difficult to modify.

To overcome this problem, the ISO has adopted a layered approach for the reference model. The complete communication subsystem is broken down into a number of layers each of which performs a well-defined function. Conceptually, these layers can be considered as performing one of two generic functions, network-dependent functions and application-oriented functions. This in turn gives rise to three distinct operational environments:

1. The **network environment** is concerned with the protocols and standards relating to the different types of underlying data communication networks.
2. The **OSI environment** embraces the network environment and adds additional application-oriented protocols and standards to allow end systems (computers) to communicate with one another in an open way.
3. The **real systems environment** builds on the OSI environment and is concerned with a manufacturers own proprietary software and services which have been developed to perform a particular distributed information processing task.

This is shown in diagrammatic form in Fig. 41.2.

Both the network-dependent and application-oriented (network-independent) components of the OSI model are implemented as a number of layers. The boundaries between each layer and

Samsung Exhibit 1022, Page 80 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

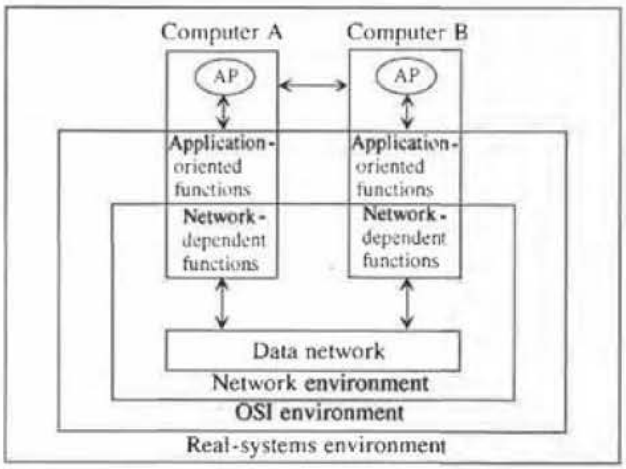


FIGURE 41.2 Operational environments.

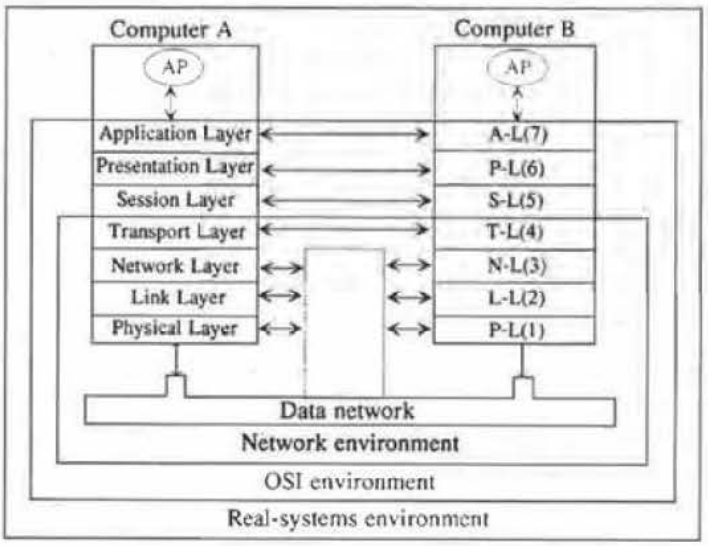


FIGURE 41.3 Overall structure of the ISO reference model.

the functions performed by each layer have been selected on the basis of experience gained during earlier standardization activity.

Each layer performs a well-defined function in the context of the overall communication subsystem. It operates according to a defined *protocol* by exchanging messages, both user data and additional control information, with a corresponding *peer* layer in a remote system. Each layer has a well-defined interface between itself and the layer immediately above and below. Consequently, the implementation of a particular protocol layer is independent of all other layers.

The logical structure of the ISO reference model is made up of seven protocol layers, as shown in Fig. 41.3. The three lowest layers (1–3) are network dependent and are concerned with the protocols associated with the data communication network being used to link the two communicating computers. In contrast, the three upper layers (5–7) are application oriented and are concerned with the

Samsung Exhibit 1022, Page 81 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

protocols that allow two end user application processes to interact with each other, normally through a range of services offered by the local operating system. The intermediate transport layer (4) masks the upper application-oriented layers from the detailed operation of the lower network-dependent layers. Essentially, it builds on the services provided by the latter to provide the application-oriented layers with a network-independent message interchange service.

The function of each layer is specified formally as a protocol that defines the set of rules and conventions used by the layer to communicate with a similar peer layer in another (remote) system. Each layer provides a defined set of services to the layer immediately above. It also uses the services provided by the layer immediately below it to transport the message units associated with the protocol to the remote peer layer. For example, the transport layer provides a network-independent message transport service to the session layer above it and uses the service provided by the network layer below it to transfer the set of message units associated with the transport protocol to a peer transport layer in another system. Conceptually, therefore, each layer communicates with a similar peer layer in a remote system according to a defined protocol. However, in practice the resulting protocol message units of the layer are passed by means of the services provided by the next lower layer. The basic functions of each layer are summarized in Fig. 41.4.

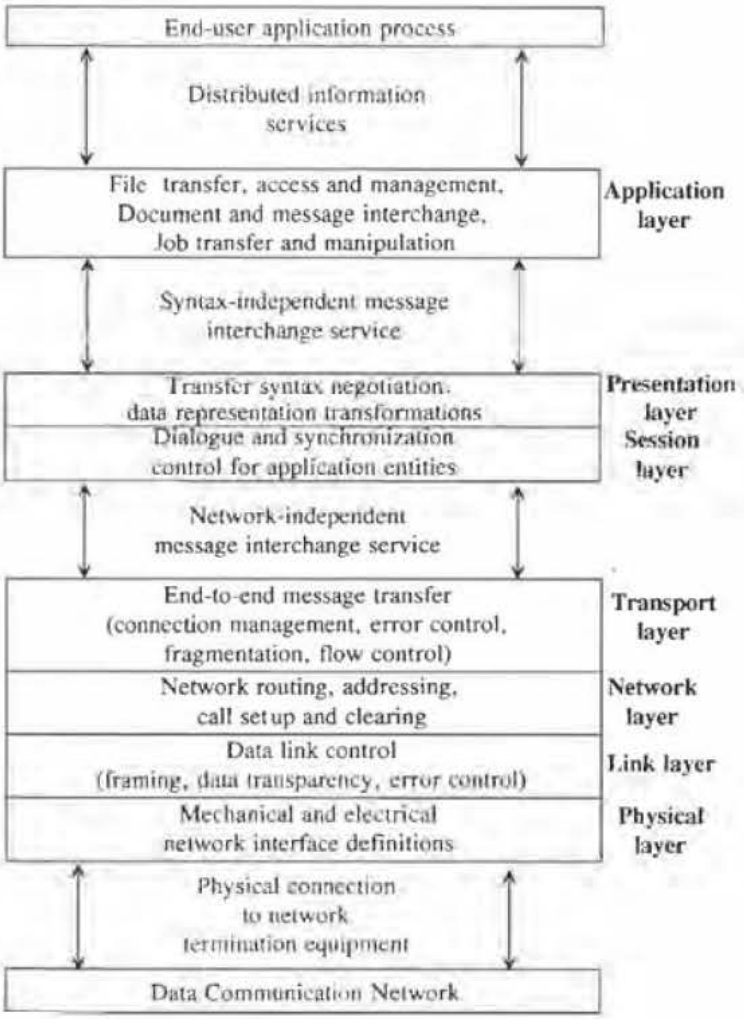


FIGURE 41.4 Protocol layer summary.

Samsung Exhibit 1022, Page 82 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The Application-Oriented Layers

The Application Layer

The application layer provides the user interface, normally an application program/process, to a range of networkwide distributed information services. These include file transfer access and management, as well as general document and message interchange services such as electronic mail. A number of standard protocols are either available or are being developed for these and other types of service.

Access to application services is normally achieved through a defined set of primitives, each with associated parameters, which are supported by the local operating system. The access primitives are the same as other operating system calls (as used for access to, say, a local file system) and result in an appropriate operating system procedure (process) being activated. These operating system procedures use the communication subsystem (software and hardware) as if it is a local device, similar to a disk controller, for example. The detailed operation and implementation of the communication subsystem is thus transparent to the (user) application process. When the application process making the call is rescheduled (run), one (or more) status parameters are returned indicating the success (or otherwise) of the network transaction that has been attempted.

In addition to information transfer, the application layer provides such services as:

- Identification of the intended communication partner(s) by name or by address
- Determination of the current availability of an intended communication partner
- Establishment of authority to communicate
- Agreement on privacy (encryption) mechanisms
- Authentication of an intended communication partner
- Selection of the dialogue discipline, including the initiation and release procedures
- Agreement on responsibility for error recovery
- Identification of constraints on data syntax (character sets, data structures, etc.)

The Presentation Layer

The *presentation layer* is concerned with the representation (syntax) of data during transfer between two communicating application processes. To achieve true open systems interconnection, a number of common *abstract data syntax* forms have been defined for use by application processes together with associated *transfer* (or *concrete*) *syntaxes*. The presentation layer negotiates and selects the appropriate transfer syntax(es) to be used during a transaction so that the syntax (structure) of the messages being exchanged between two application entities is maintained. Then, if this form of representation is different from the internal abstract form, the presentation entity performs the necessary conversion.

To illustrate the services provided by the presentation layer, consider a telephone conversation between a French speaking person and a Spanish speaking person. Assume each uses an interpreter and that the only language understood by both interpreters is English. Each interpreter must translate from their local language to English, and vice versa. The two correspondents are thus analogous to two application processes with the two interpreters representing presentation layer entities. French and Spanish are the local syntaxes and English the transfer or concrete syntax. Note that there must be a universally understood language which must be defined to allow the agreed transfer language (syntax) to be negotiated. Also note that the interpreters do not necessarily understand the meaning (semantics) of the conversation.

Another function of the presentation layer is concerned with data security. In some applications, data sent by an application is first encrypted (enciphered) using a *key*, which is (hopefully) known only by the intended recipient presentation layer. The latter decrypts (deciphers) any received data using the corresponding key before passing it on to the intended recipient.

Samsung Exhibit 1022, Page 83 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The Session Layer

The *session layer* provides the means that enables two application layer protocol entities to organize and synchronize their dialogue and manage their data exchange. It is thus responsible for setting up (and clearing) a communication (dialogue) channel between two communicating application layer protocol entities (presentation layer protocol entities in practice) for the duration of the complete network transaction. A number of optional services are provided, including the following:

- Interaction management: The data exchange associated with a dialogue may be duplex or half-duplex. In the latter case it provides facilities for controlling the exchange of data (dialogue units) in a synchronized way.
- Synchronization: For lengthy network transactions, the user (through the services provided by the session layer) may choose periodically to establish synchronization points associated with the transfer. Then, should a fault develop during a transaction, the dialogue may be restarted at an agreed (earlier) synchronization point.
- Exception reporting: Nonrecoverable exceptions arising during a transaction can be signaled to the application layer by the session layer.

The Transport Layer

The *transport layer* acts as the interface between the higher application-oriented layers and the underlying network-dependent protocol layers. It provides the session layer with a message transfer facility that is independent of the underlying network type. By providing the session layer with a defined set of message transfer facilities, the transport layer hides the detailed operation of the underlying network from the session layer.

The transport layer offers a number of **classes of service** which cater for the varying *quality of service* (*QOS*) provided by different types of network.

There are five classes of service ranging from class 0, which provides only the basic functions needed for connection establishment and data transfer, to class 4, which provides full error control and flow control procedures. As an example, class 0 may be selected for use with a packet-switched data network (PSDN), whereas class 4 may be used with a local area network (LAN) providing a best-try service; that is, if errors are detected in a frame then the frame is simply discarded.

The Network-Dependent Layers

As the lowest three layers of the ISO reference model are network dependent, their detailed operation varies from one network type to another. In general, however, the *network layer* is responsible for establishing and clearing a networkwide connection between two transport layer protocol entities. It includes such facilities as network routing (addressing) and, in some instances, flow control across the computer-to-network interface. In the case of internetworking it provides various harmonizing functions between the interconnected networks.

The *link layer* builds on the physical connection provided by the particular network to provide the network layer with a reliable information transfer facility. It is thus responsible for such functions as error detection and, in the event of transmission errors, the retransmission of messages. Normally, two types of service are provided:

1. *Connectionless* treats each information frame as a self-contained entity that is transferred using a best-try approach.
2. *Connection oriented* endeavors to provide an error-free information transfer facility.

Finally, the *physical layer* is concerned with the physical and electrical interfaces between the user equipment and the network terminating equipment. It provides the link layer with a means of transmitting a serial bit stream between the two equipments.

Samsung Exhibit 1022, Page 84 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

41.4 Open System Standards

The ISO reference model has been formulated simply as a template for the structure of a communication subsystem on which standards activities associated with each layer may be based. It is not intended that there should be a single standard protocol associated with each layer. Rather, a set of standards is associated with each layer, each offering different levels of functionality. Then, for a specific open systems interconnection environment, such as that linking numerous computer-based systems in a fully automated manufacturing plant, a selected set of standards is defined for use by all systems in that environment.

The three major international bodies actively producing standards for computer communications are the ISO, the American Institute of Electrical and Electronic Engineers (IEEE) and the International Telegraph and Telephone Consultative Committee (CCITT). Essentially, ISO and IEEE produce standards for use by computer manufacturers, whereas CCITT defines standards for connecting equipment to the different types of national and international public networks. As the degree of overlap between the computer and telecommunications industries increases, however, there is an increasing level of cooperation and commonality between the standards produced by these organizations.

In addition, prior to and concurrently with ISO standards activity, the U.S. Department of Defense has for many years funded research into computer communications and networking through its Defense Advanced Research Projects Agency (DARPA). As part of this research, the computer networks associated with a large number of universities and other research establishments were linked to those of DARPA. The resulting internetwork, known as ARPANET, has been extended to incorporate internets developed by other government agencies. The combined internet is now known simply as the *Internet*.

The protocol suite used with the Internet is known as **transmission control protocol/internet protocol (TCP/IP)**. It includes both network-oriented protocols and application support protocols. Because TCP/IP is in widespread use with an existing internet, many of the TCP/IP protocols have been used as the basis for ISO standards. Moreover, since all of the protocol specifications associated with TCP/IP are in the public domain, and hence no licence fees are payable, they have been used extensively by commercial and public authorities for creating open system networking environments. In practice, therefore, there are two major open system (vendor-independent) standards: the TCP/IP protocol suite and those based on the evolving ISO standards.

Figure 41.5 shows some of the standards associated with the TCP/IP protocol suite. As can be seen, since TCP/IP has developed concurrently with the ISO initiative, it does not contain specific protocols relating to all of the seven layers in the OSI model.

Moreover, the specification methodology used for the TCP/IP protocols differs from that used for the ISO standards. Nevertheless, most of the functionality associated with the ISO layers is embedded in the TCP/IP suite. A range of standards has been defined by the ISO/CCITT and a selection of these is shown in Fig. 41.5. Collectively they enable the administrative authority that is establishing the open system environment to select the most suitable set of standards for the application. The resulting protocol suite is known as the *open system interconnection profile*. A number of such profiles have now been defined, including: TOP, a protocol set for use in technical and office environments; MAP, for use in manufacturing automation; U.S. and U.K. GOSIP for use in U.S. and U.K. government projects, respectively; and a similar suite used in Europe known as the CEN functional standards. The latter has been defined by the Standards Promotion and Application Group (SPAG), a group of 12 European companies.

As Fig. 41.6 shows, the lower three layers vary for different network types. CCITT has defined V, X, and I series standards for use with public-carrier networks. The V series is for use with the existing switched telephone network (PSTN), the X series for use with existing switched data networks (PSDN), and the I series for use with the integrated services digital networks (ISDN). Those produced by ISO/IEEE for use with local area networks are known collectively as the 802 (IEEE) or 8802 (ISO) series.

Samsung Exhibit 1022, Page 85 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

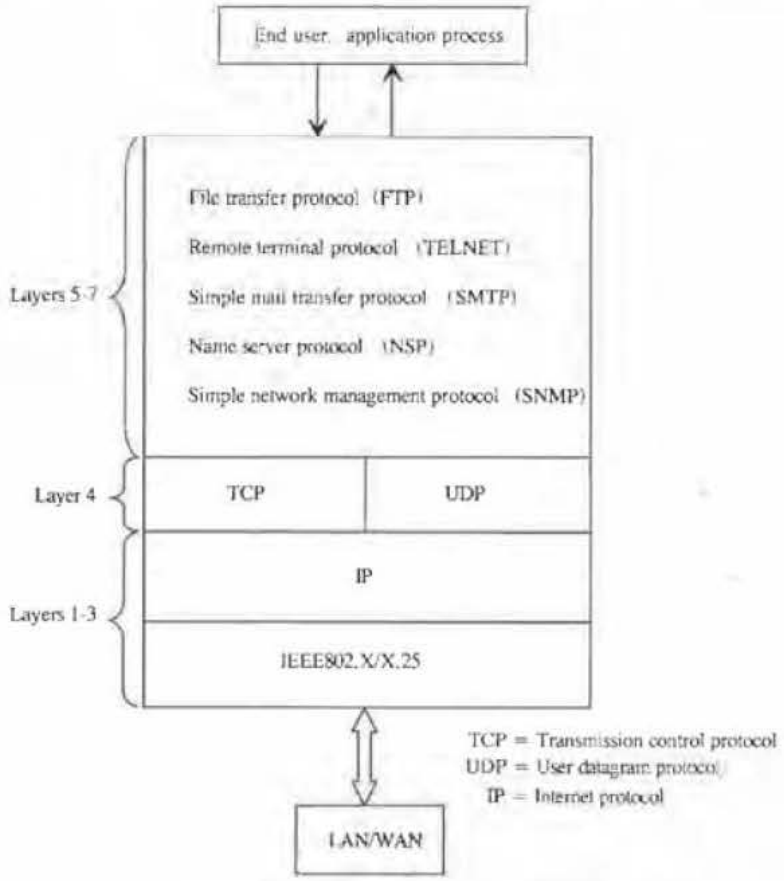


FIGURE 41.5 TCP/IP protocol suite.

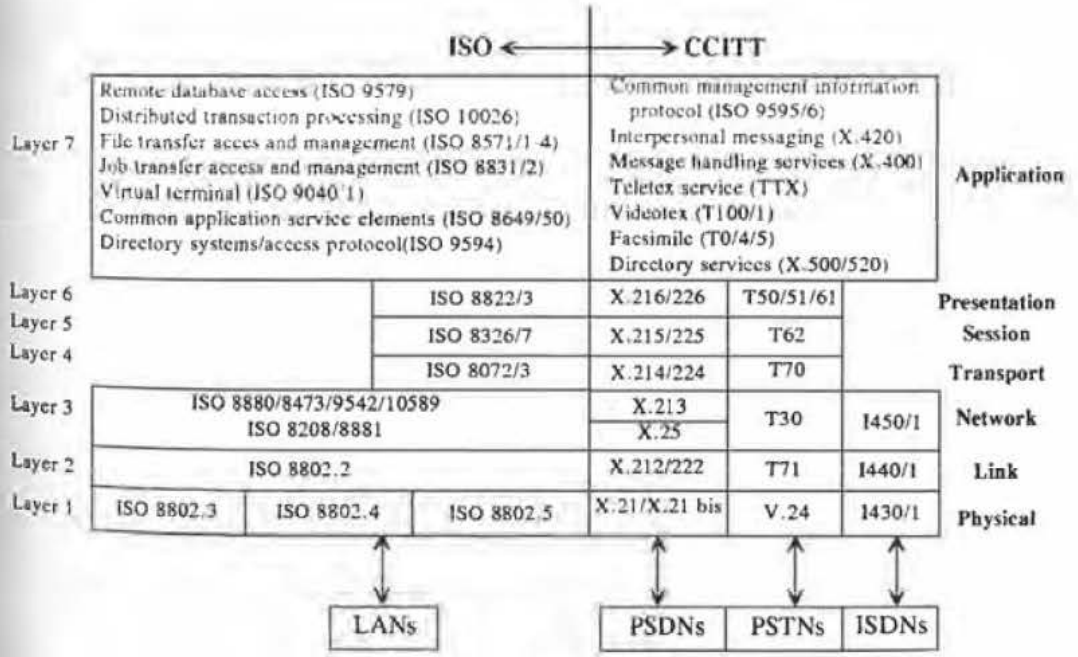


FIGURE 41.6 Standards summary.

Samsung Exhibit 1022, Page 86 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

41.5 Summary

This chapter has reviewed the requirements for the communications subsystem in each of a set of interconnected computers that enables them to communicate in an open way to perform various distributed application functions. The philosophy behind the structure of the ISO reference model for open systems interconnection has been presented and a description of the functionality of the seven layers that make up the reference model described. Finally, a selection of the ISO/CCITT standards that have been defined have been identified.

Defining Terms

Internetwork: A collection of different network types that are linked together.

Local area network (LAN): Network used to link computers that are distributed around a single site.

Open system interconnection (OSI): A standard set of communication protocols that are independent of a specific manufacturer.

Transmission control protocol/Internet protocol (TCP/IP): The standard set of communication protocols that are used by each computer that is connected to the Internet.

Wide area network (WAN): Network used to link computers that are situated at different sites.

Further Information

Further coverage of the material covered in this chapter can be found in:

Halsall, F. 1996. *Data Communications, Computer Networks and Open Systems, 4th ed.*, Addison Wesley, Reading, MA.

Samsung Exhibit 1022, Page 87 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

once the station has determined its packet to be involved in a process of collision, the station will immediately abort transmission. In this manner, colliding stations will occupy the channel with their colliding transmissions only for a limited period of time, the collision detection time $T(\mathrm{CD})$. Clearly, if $T(\mathrm{CD}) < T(P)$, where $T(P)$ denotes the transmission time of a packet (MAC frame) across the medium, the CSMA/CD operation leads to improved delay-throughput performance over that of a CSMA operation.

The ethernet LAN developed by Xerox Corporation is a bus-based network operating at a channel data rate of 10 Mb/s and using a 1-persistent CSMA/CD medium access control algorithm. The physical, MAC and link layers of such a CSMA/CD network are defined by the IEEE 802.3 Standard (and a corresponding ISO standard). Ethernet nets can employ different media types: twisted pair, coaxial cable, fiber optic line, as well as radio links (in the case of an ethernet-based wireless LAN system). The configuration of the ethernet LAN is that of a logical bus so that a frame transmission by a station propagates across the bus medium to all other stations. In turn, the physical layout of the medium can assume a bus or a star topology. Under the latter configuration, all stations are connected to a central hub node at which point the access lines are connected by a reflecting repeater module so that a transmission received from an access line is repeated into all other lines. Typical ethernet layouts are limited in their geographical span to an overall distance of about 500–2000 m.

The delay-throughput efficiency of the CSMA/CD network operation is determined by the acquisition time delay $t(a)$ and index a, as for the CSMA network. In addition, the network efficiency also depends on the CD time $T(\mathrm{CD})$. It is noted that $T(\mathrm{CD})$ can be as long as twice the propagation delay across the overall span of the bus medium, plus the time required by the station to establish the occurrence of a collision. To ensure that a collision is reliably detected, the power levels of received packets must be sufficiently high. This also serves as a factor limiting the length of the bus and of the distance at which repeaters must be placed. As a result, the network stations are attached to several ethernet *segments*, each assuming a sufficiently short span. The different ethernet segments are interconnected by gateways or routers. The latter act as store-and-forward switches which serve to isolate the CSMA/CD multiaccess operation of each segment.

As for the random-access mechanisms previously discussed, burst loads applied to the CSMA/CD network can cause large delay-throughput degradations. In supporting application streams which require limited packet delay jitters, it is thus required to prevent the bus from being excessively loaded. Many implementations thus plan the net's offered traffic loading levels to be no higher than 40% (or 4 Mb/s). Also note that as for other random access schemes, large loading variations can induce an unstable behavior. A flow control mechanism must then be employed to regulate the maximum loading level of the network.

When shorter bus spans are used, or when shorter access link distances are employed in a star configuration, it is possible to operate this network at higher channel data rates. Under such conditions, an ethernet operation at a data rate of 100 Mb/s (or higher) has been implemented.

46.6 Multiple-Access Schemes for Wireless Networks

Under a cellular wireless network architecture, the geographical area is divided into cells. Each cell contains a central base station. The mobile terminals communicate with the base station controlling the cell in which they reside. The terminals use the reverse traffic and signalling channel(s) to transmit their messages to the base station. The base station multiplexes the messages it wishes to send to the cell's mobile terminals across the forward traffic and signalling channels.

First-generation cellular wireless networks are designed to carry voice connections employing a circuit switching method. Analog communications signals are used to transport the voice information. The underlying signalling subnetwork is used to carry the connection setup, termination, and handover signalling messages. The voice circuits are allocated through the use of a reservation based demand-assigned/FDMA scheme. A ready mobile uses an allocated signalling channel for its handset to transmit to its cell's base station a request message for the establishment of a voice

Samsung Exhibit 1022, Page 88 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

circuit. If a frequency channel is available for the allocation of such a circuit (traffic channel), the base station will make this allocation by signalling the requesting handset.

Second-generation cellular wireless networks use digital communications channels. A circuit-switching method is still employed, with the primary service providing for the accommodation of voice connections. Circuits are formed across the shared radio medium in each cell by using either a TDMA-access control scheme (through the periodic allocation of time slots to form an established circuit, as performed by the European GSM and the U.S. IS-54 standards) or by employing a CDMA procedure (through the allocation of a code sequence to a connection's circuit, as carried out by the IS-95 standard). A signalling subnetwork is established. Reverse signalling channels (RSCs) are multiple-access channels which are used by the mobile terminals to transmit to their cell's base station their channel-request packets (for a mobile originating call), as well as for the transmission of paging response packets (by those mobiles which terminate calls). Forward signalling channels (FSCs) are multiplexing channels configured for the transmission of packets from the base station to the mobiles. Such packets include channel-allocation messages (which are sent in response to received channel request packets) and paging packets (which are broadcasted in the underlying location area in which the destination mobile may reside).

For the reverse signalling channels, a random access algorithm such as the ALOHA scheme is frequently employed. For TDMA systems, time circuit(s) are allocated for the random access transmission of signalling packets. For CDMA systems, codes are allocated for signalling channels; each code channel is time shared through the use of a random access scheme (employing, for example, a slotted or unslotted ALOHA multiple-access algorithm).

Paging and channel allocation packets are multiplexed (in a time-shared fashion) across the forward signalling channels. To reduce battery consumption at the mobile terminals, a slotted mode operation can be invoked. In this case, the cell mobiles are divided into groups, and paging messages destined to a mobile belonging to a certain group are transmitted within that group's allocated channels (time slots) [Rubin and Choi, 1996]. In this manner, an idle mobile handset needs to be activated for the purpose of listening to its FSC only during its group's allocated slots.

Under the IEEE 802.11 protocol, terminals access the base station by sharing the radio channel through the use of a CSMA/CA (collision avoidance) scheme.

Third-generation cellular wireless networks are planned to provide for the support of both voice and data services. These networks employ packet-switching principles. Many of the multiple-access schemes described in the previous sections can be employed to provide for the sharing of the mobile-to-base-station multiple-access reverse communications channels. In particular, random-access, polling (or implicit polling) and reservation protocols can be effectively implemented. For example, the following versions of a reservation method have been investigated. Under the packet reservation multiple access (PRMA) [Goodman et al., 1989] or random-access burst reservation procedure, a random-access mechanism is used for the mobile to reserve a time circuit for the transmission of a burst (which consists of a number of consecutively generated packets). For voice bursts, a random-access algorithm is used to govern the transmission of the first packet of the burst across the shared channel (by randomly selecting an unused slot, noting that, in each frame, the base station notifies all terminals as to which slots are unoccupied). If this packet's transmission is successful, its terminal keeps the captured slot's position in each subsequent frame until the termination of the burst's activity. Otherwise, the voice packet is discarded and the next voice packet is transmitted in the same manner. In selecting the parameters and capacity levels of such a multiaccess network, it is necessary to ensure that connections receive acceptable throughput levels and that the packet discard probability is not higher than a prescribed level ensuring an acceptable voice quality performance [Rubin and Shambayati, 1995].

For connection-oriented packet-switching network implementations which provide also data packet transport, including wireless ATM networks, it is necessary to avoid frequent occurrences of packet discards, to reduce the rate of packet retransmissions and to lower the chances for packet reordering procedures needed to guarantee the ordered delivery of packets. Reverse and forward

Samsung Exhibit 1022, Page 89 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

signalling channels are established to set up the virtual-circuit connection. Channel resources can then be allocated to established connections in accordance with the statistical features of such connections. For example, real-time connections can be accommodated through the periodic allocation of time slots (or frequency/code resources) whereas burst sources can be supported by the allocation of such resources only for the limited duration of a burst activity. A mechanism must then be employed to identify the start and end times of burst activities. For example, the signalling channels can be used by the mobiles to signal the start of burst activity; whereas the in-band channel is used to detect the end of activity (directly and/or through the use of tagging marks).

46.7 Multiple Access Methods for Spatial-Reuse Ultra-High-Speed Optical Communications Networks

As we have seen above, token-ring and FDDI LAN's use token-passing methods to access a shared medium. Furthermore, these networks use a *source removal* procedure so that each station is responsible for removing its own transmitted frames from the ring. In this manner, each transmitted frame circulates the ring and is then removed by its source station. As a result, the throughput capacity of such networks is limited to a value which is not higher than the ring's channel data rate.

In turn, as observed in Sec. 46.4 (in connection with implicit polling-based multiple-access networks), the bandwidth utilization of shared medium networks, with particular applications to local and metropolitan area networks operating at high speed using generally fiber optic links, can be significantly upgraded through the employment of spatial-reuse methods. For example, consider a ring network which consists of two counter-rotating unidirectional fiber optic ring topologies. Each terminal (station) is connected to both rings through ring interface units. The bandwidth resources of the rings are shared among the active stations. Assume now that a *destination removal* method is employed. Thus, when a frame reaches its destination node (RIU), it is removed from the ring by the latter. The network communications resources occupied by this frame (such as a time slot) can then be made available to source nodes located downstream to the removing station (as well as to the removing station itself).

A source station which has a frame to transmit selects the ring that offers the shortest path to the destination node. In this manner, the length of a path will be no longer than 1/2 the ring's length. If we assume a uniform traffic matrix (so that traffic flows between the network terminal nodes are symmetrically distributed), the average path length will be equal to 1/4 the ring's length. As a result, using such a destination removal technique, we conclude that an average of four source-destination flows will be carried simultaneously in time by each one of the two rings. Therefore, the throughput capacity of such a spatial reuse network can reach a level which is equal to 400% the ring's channel data rate, for each one of the two rings.

Spatial reuse ring networks can be used to carry traffic on a circuit-switched or packet-switched (connectionless or connection-oriented) basis. For packet-switched network implementations, the following two distributed implicit polling multiple-access methods have been used. Under a slotted ring operation, time slots are generated and circulated across the ring. Each time slot contains a header which identifies it as either empty or busy. A station which senses an empty time slot can mark it as busy and insert a segment (a MAC frame, which for example can carry an ATM cell) into the slot. The destination station will remove the segment from its slot, and will designate the slot as idle so that it can be reused by itself or other stations. Under a buffer-insertion ring multiple-access scheme, each station inserts a buffer into the ring. A busy station will defer the transmission of its packet to packets currently being received from an upstream station. In turn, when its buffer is detected to be empty, the station will insert its segment into the ring. The buffer capacity is set equal to a maximum size packet so that a packet received by a station while it is in a process of transmission of its own packet can be stored in the station's inserted buffer. Each packet is removed from the ring by its destination node.

As previously noted, the DQDB MAN is a slotted dual-bus network which employs active station interfaces. To achieve spatial bandwidth reuse gains of the bus' bandwidth resources, frame removal

Samsung Exhibit 1022, Page 90 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

47

Routing and Flow Control

47.1 Introduction 650
47.2 Connection-Oriented and Connectionless Protocols, Services, and Networks 651
47.3 Routing in Datagram Networks 652
47.4 Routing in Virtual Circuit Switched Networks 655
47.5 Hierarchical Routing 655
47.6 Flow Control in Datagram Networks 656
47.7 Flow Control in Virtual Circuit Switched Networks 657

Rene L. Cruz
University of California, San Diego

47.1 Introduction

There are a large number of issues relating to routing and flow control, and we cannot hope to discuss the subject comprehensively in this short chapter. Instead, we outline generic concepts and highlight some important issues from a high-level perspective. We limit our discussion to the context of **packet switched networks.**

Routing and flow control are concerned with how information packets are transported from their source to their destination. Specifically, routing is concerned with *where* packets travel, that is, the paths traveled by packets through the network. Flow control is concerned with *when* packets travel, that is, the timing and rate of packet transmissions.

The behavior of routing and flow control can be characterized by a number of interrelated performance measures. These performance measures can be categorized into two groups: user metrics and network metrics. User metrics characterize performance from the perspective of a single user. Common user metrics are throughput (rate of information transfer), average delay, maximum delay, delay jitter (variations in delay), and packet loss probability. Network metrics characterize performance from the perspective of the network as a whole, and all users of it. Common network metrics are network throughput, number of users that can be supported, average delay (averaged over all users), and buffer requirements. Fairness issues may also arise from the perspective of the network.

The goals of routing and flow control are to balance these different performance measures. Optimization of user metrics are often at the expense of network metrics, and vice versa. For example, if the number of users allowed to access the network is too large, the quality of service seen by some users might fall below acceptable levels.

In reality, it is often even difficult to optimize a single performance measure, due to the uncertainty of demand placed on the network. Statistical models can be used, but the appropriate parameters of the models are often unknown. Even with an accurate model of demand, the problem of predicting

0-8493-8349-8/97/$0.00+$.50

© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 91 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

performance for a given routing or flow control strategy is often intractable. For this reason, routing and flow control algorithms are typically based on several simplifying assumptions, approximations, and heuristics.

Before we discuss routing and flow control in more detail, we begin by classifying different modes of operation common in packet switched networks.

47.2 Connection-Oriented and Connectionless Protocols, Services, and Networks

In a connection-oriented protocol or service between two network terminals, the two terminals interact with one another prior to the transport of data, to set up a *connection* between them. The interaction prior to the exchange of data is required, for example, to identify the terminals to each other and to define the rules governing the exchange of data. For example, users of the telephone network engage in a connection-oriented protocol in the sense that a phone number must first be dialed, and typically the parties involved must identify themselves to each other before the essential conversation can begin. The transport control protocol (TCP) used on the Internet is an example of a connection-oriented protocol. Examples of connection-oriented services are rlogin and telnet, which can be built upon TCP. Connection-oriented protocols and services typically involve the exchange of data in both directions. In addition, connection-oriented protocols are frequently used to provide reliable transport of data packets between network endpoints, whereby the network endpoints continue to interact during the data exchange phase in order to manage retransmission of lost or errored packets.

In *connectionless* protocols or services between two network terminals, no interaction between the terminals takes place prior to the exchange of data. For example, when a letter is mailed over the postal network, a connectionless protocol is used in the sense that the recipient of the letter need not interact with the sender ahead of time. An example of a connectionless protocol is the current Internet protocol (IP) used to transport packets on the Internet. Electronic mail provides a connectionless service, in the same sense that the postal network provides a connectionless service. Because of the lack of interaction between terminals engaging in a connectionless protocol, retransmission of lost or errored packets is not possible, and hence connectionless protocols are not considered intrinsically reliable. On the other hand, connectionless protocols are often appropriate in environments where lost or errored packets are sufficiently rare.

A connection-oriented *network* is a network specifically designed to support connection-oriented protocols. When a connection request is made to a connection-oriented network, the network establishes a path between the source and destination terminal. The switching nodes are then configured to route packets from the connection along the established path. Each switching node is aware of the connections that pass through it and may reserve resources (e.g., buffers and bandwidth) on a per-connection basis. A packet switched network that is connection oriented is also referred to as a **virtual circuit switched network.** Asynchronous transfer mode (ATM) networks are an example of virtual circuit switched networks.

A connectionless *network* is a network designed to support connectionless protocols. Each node of a connectionless network typically knows how to route packets to each possible destination, since there is no prior coordination between terminals before they communicate. Connectionless networks are sometimes referred to as **datagram networks.**

The terminology as defined is not standard, and indeed there does not appear to be widespread agreement on terminology. We have made a distinction between a connection-oriented protocol and a connection-oriented network, and a distinction between a connectionless protocol and a connectionless network. The reason for these distinctions is because of the many ways in which protocols and networks can be layered. In particular, connection-oriented protocols can be built upon connectionless networks and protocols, and connectionless protocols can be built upon

Samsung Exhibit 1022, Page 92 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

connection-oriented networks and protocols. For example, TCP is in widespread use on end hosts in the Internet, but the Internet itself is inherently a connectionless network which utilizes IP. Another example that has received attention recently is in using an ATM network (connection oriented) as a lower layer transport mechanism to support the IP protocol between Internet nodes. This may be considered somewhat odd, since IP will in turn often support the connection-oriented TCP.

Packet switched networks can be further classified as lossy or lossless according to whether or not network nodes will sometimes be forced to drop (destroy) packets without forwarding them to their intended destination. Connectionless networks are inherently lossy, since the lack of coordination of data sources creates the possibility that buffers inside the network will overflow. Connection-oriented networks can be designed to be lossless. On the other hand, connection-oriented networks can also be lossy, depending on if and how flow control is exercised.

47.3 Routing in Datagram Networks

We will first confine our discussion on routing to the context of connectionless networks. A common approach to routing in connectionless networks is for each node in the network to have enough intelligence to route packets to every possible destination in the network.

A common approach to route computation is *shortest path routing*. Each link in the network is assigned a weight, and the length of a given path is defined as the sum of the link weights along the path. The weight of a link can reflect the desirability of using that link and might be based on reliability or queueing delay estimates on the link, for example. However, another common choice for link weights is unity, and shortest paths in this case correspond to minimum hop paths.

We now briefly review some popular algorithms for computing shortest paths. Suppose there are n nodes in the network, labeled $1, 2, \ldots, n$. The label (i, j) is assigned to a link that connects node i to node j. The links are assumed to be oriented, that is, link (i, j) is distinct from (j, i). We assume that if there is a link (i, j) in the network, then there is also a link (j, i) in the network. We say that node i is a neighbor of node j if the link (i, j) exists. Let $N(i)$ be the set of nodes for which node i is a neighbor. Let w_{ij} denote the weight assigned to link (i, j), and define $w_{ij} = +\infty$ if node i is not a neighbor of node j. We assume that $w_{ij} > 0$ for all i and j. Formally, a *path* from node i_1 to node i_k is a sequence of nodes $(i_1, i_2, \ldots, i_k)$. The length of such a path is the sum of the weights of links between successive nodes, and the path is said to have $k - 1$ hops. We assume that the network is *strongly connected*, that is, there exists a path between all ordered pairs of nodes with finite length.

Consider the problem of finding shortest (i.e., minimum length) paths from each node to a given destination node. Since each node in the network is a potential destination, this problem will be solved separately for each destination. Without loss of generality, assume that the destination node is node 1. The distance from a node i to the destination node is defined to be the length of the shortest path from node i to the destination node, and is denoted as D_i. Define $D_1 = 0$.

Consider a shortest path from a node i to the destination: $(i, j, k, \ldots, 1)$. Note that the subpath that begins at node j, $(j, k, \ldots, 1)$, must be a shortest path from node j to the destination, for otherwise there exists a path from node i to the destination that is shorter than $(i, j, k, \ldots, 1)$. Thus, this subpath has length D_j, and $D_i = w_{ij} + D_j$. Furthermore, since D_i is the length of a shortest path, we have $D_i \leq w_{ik} + D_k$ for any k. Thus, we have $D_1 = 0$ and

$$D_i = \min\{w_{ij} + D_j : 1 \leq j \leq n\}, \qquad 1 < i \leq n$$

This is known as *Bellman's equation*, and forms the basis of an optimization technique known as dynamic programming. Let $\boldsymbol{D} = (D_1, D_2, \ldots, D_n)$ denote the vector of shortest path lengths, and let $\boldsymbol{f}$ denote the vector valued mapping in Bellman's equation. With these definitions, Bellman's equation says that $\boldsymbol{D}$ is a fixed point of the mapping $\boldsymbol{f}$, that is, $\boldsymbol{D} = \boldsymbol{f}(\boldsymbol{D})$. It can be shown that under our assumption of positive link weights, there is only one solution $\boldsymbol{x}$ to the equation $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})$, that is, $\boldsymbol{D}$ is the unique solution to Bellman's equation.

Samsung Exhibit 1022, Page 93 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The *Bellman–Ford algorithm* for shortest paths finds the solution to Bellman's equation iteratively. Specifically, let $\boldsymbol{D}^0 = (0, \infty, \infty, \ldots, \infty)$. In the first iteration of the Bellman–Ford algorithm, $\boldsymbol{D}^1 = f(\boldsymbol{D}^0)$ is computed. In the mth iteration, $\boldsymbol{D}^m = f(\boldsymbol{D}^{m-1})$ is computed. It can be shown by induction on m that $\boldsymbol{D}^m$ is the vector of shortest path lengths, where the paths are constrained to have at most m hops. Since link weights are positive, any shortest path visits a node at most once, and hence any shortest path has at most $n-1$ hops. Thus, it follows that the Bellman–Ford algorithm converges within $n-1$ iterations. In other words, we have $f(\boldsymbol{D}^{n-1}) = \boldsymbol{D}^{n-1} = \boldsymbol{D}$. The shortest paths can be generated from $\boldsymbol{D}$ by examining the argument of the minimum in Bellman's equation. Specifically, for each i, let $\text{Next}(i) = j$, where j is an argument of the minimum in the equation for D_i, that is, $D_i = w_{ij} + D_j$. With this definition, $\text{Next}(i)$ is clearly the first node along a shortest path from node i to the destination. The worst case time complexity of the Bellman–Ford algorithm is easily seen to be $\mathcal{O}(n^3)$.

The Bellman–Ford shortest path algorithm is sometimes called a *distance vector* routing algorithm. The algorithm lends itself easily to a distributed, asynchronous implementation that forms the basis for many commonly used routing algorithms. In such an implementation, each node i is responsible for computing the distance from itself to the destination D_i. Each node initially has an estimate of D_i. These initial estimates may be arbitrary, except the destination itself always knows the correct distance from itself to itself, that is, node 1 knows that $D_1 = 0$. Each node i sends, or advertises, to its neighbors the estimate of D_i that it currently has. This advertisement is either done periodically or whenever the estimate changes. Whenever a node receives an estimate from a neighbor, it updates its own estimate in accordance with Bellman's equation. Specifically, if $D_i(t)$ is the estimate of D_i at time t and if $D_j(t-\tau_j)$ is the estimate of D_j from node j that was most recently advertised to node i, then $D_i(t) = \min\{w_{ij} + D_j(t-\tau_j) : j \in N(i)\}$. Note that this minimum can be taken over only nodes j which are neighbors of node i, since $w_{ij} = \infty$ if node i is not a neighbor of node j. It can be shown that each node's estimate converges to the actual distance to the destination under essentially arbitrary assumptions regarding propagation delays of advertisements, and delays associated with computing estimates [Bertsekas and Gallager, 1992].

In the preceding discussion we assumed that the link weights were constant. In practice, however, the link weights may change with time. For example, this may be as a result of link failures, or changing traffic characteristics on the links. If the link weights change slowly enough, the distributed asynchronous Bellman–Ford algorithm will track the shortest paths correctly as the link weights change. However, if the link weights change fast enough so that the algorithm does not have a chance to converge, problems may arise. For example, if link weights change quickly in response to changing traffic characteristics, it may be possible for packets to travel around in loops indefinitely. In such a scenario, the link weights oscillate as the traffic pattern on the links change due to oscillating estimates of shortest paths [Bertsekas and Gallager, 1992].

Another popular approach to finding shortest paths is for each node to independently compute the shortest paths. Each node i broadcasts the link weight w_{ij} for all outgoing links (i, j) to all other nodes. Each node then has complete information about the network and can compute shortest paths independently. Networks that use this approach are sometimes said to use *link state* routing. Each node could use the Bellman–Ford algorithm as described to compute shortest paths. However, with link state routing it is common to use *Dijkstra's algorithm* for shortest paths, described next.

Dijkstra's algorithm for finding shortest paths first finds the closest node to the destination node (i.e., the node i other than the destination node with the smallest value of D_i), then the second closest node, then the third closest node, and so on. Let us assume again without loss of generality that node 1 is the destination node. Dijkstra's algorithm is summarized as follows:

Initialization: $m \leftarrow 1$, $P_m \leftarrow \{1\}$, $D_1 \leftarrow 0$.
For all i not in P_m:
    $X_i \leftarrow w_{i1} + D_1$
    $\text{Next}(i) \leftarrow 1$

Samsung Exhibit 1022, Page 94 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Iteration:

$i^* \leftarrow \text{argmin}\{X_i : i \text{ not in } P_m\}$
$P_{m+1} \leftarrow P_m \cup \{i^*\}$
$D_{i^*} \leftarrow X_{i^*}$
For all k not in P_{m+1}:
    If $X_k > w_{ki^*} + D_{i^*}$ then
        $X_k \leftarrow w_{ki^*} + D_{i^*}$
        $\text{Next}(k) \leftarrow i^*$
$m \leftarrow m + 1$
If $m = n$, then terminate. Otherwise, execute iteration again.

The set P_m consists of the m closest nodes to the destination, including the destination. In other words, for each m we have $|P_m| = m$, and $D_i \geq D_j$ if j belongs to P_m and i does not belong to P_m. To see why Dijkstra's algorithm works, assume after executing the iteration m times that P_{m+1} is the set of $m + 1$ closest nodes, that $X_k = \min\{w_{kj} + D_j : j \in P_{m+1}\}$, and that D_j has been computed correctly for all $j \in P_{m+1}$. Clearly this is true for $m = 0$, and we now show by induction that it is true for all m. Consider the $(m + 1)$st iteration. Suppose that node i^* is a node that could be added to P_{m+1} to yield P_{m+2} with the desired property and that node k is an arbitrary node that cannot be added to P_{m+1} to yield P_{m+2} with the desired property. Note that $D_k > D_{i^*}$. By the induction hypothesis and the fact that all link weights are positive, it thus follows from Bellman's equation that $D_{i^*} = X_{i^*}$. Furthermore, we have $X_k \geq D_k > D_{i^*} = X_{i^*}$. Thus, in the $(m + 1)$st iteration, node i^* is chosen correctly to form P_{m+2} and $D_{i^*} = X_{i^*}$. Furthermore, we see that X_k is updated correctly to preserve the induction hypothesis.

It is seen that the time complexity of Dijkstra's algorithm is $\mathcal{O}(n^2)$, which is less than the worst case-time complexity of $\mathcal{O}(n^3)$ for the Bellman–Ford algorithm. The description of Dijkstra's algorithm can also easily be adapted to address the computation of shortest paths from a given source node to all possible destinations, which makes it well suited for link state routing.

Deflection routing [Maxemchuk, 1987] has been proposed for high-speed networks. With deflection routing, each node attempts to route packets along a shortest path. If a link becomes congested, however, a node will intentionally send a packet to a node which is *not* along a shortest path. Such a packet is said to be *deflected*. If the number of links incoming to a node is equal to the number of outgoing links, and the capacity of all links is the same, packet buffering can be essentially eliminated with deflection routing. It is possible for packets to circulate in loops with deflection routing. With a suitable priority structure, however, delay can be bounded. Deflection routing has also been proposed *within* high-speed packet switches [Cruz and Tsai, 1996].

In another approach to routing, sometimes called *optimal routing*, the rate of traffic flow between each possible source–destination pair is estimated. Suppose that $r(w)$ is the estimated rate (bits per second) of traffic flow for a given source–destination pair w. A number of possible paths for routing traffic between each source–destination pair are identified, say path 1 through P_w for source–destination pair w. The total flow $r(w)$ is allocated among all of the P_w paths according to a vector $(f_1, f_2, \ldots, f_{P_w})$ such that $r(w) = f_1 + f_2 + \cdots + f_{P_w}$, and such that f_p flows on path p. Typically, bifurcation is allowed, so that f_p may be positive for more than one path p. The path flow allocation for all source–destination pairs is determined simultaneously in order to minimize some cost function. Typically, the cost function is of the form $\Sigma C_e(f_e)$, where the sum is taken over all links e in the network and f_e denotes the total flow over all paths which cross link e. A common choice for $C_e(f_e)$ is $C_e(f_e) = f_e/(C - f_e)$, where C is the capacity in bits per second of link e. This would correspond to the average delay on link e if the output queue for link e were an M/M/1 queue. Nonlinear optimization techniques can be used to optimize the path flow allocations with respect to the cost function. The disadvantage of the optimal routing approach is that it is difficult to estimate the rates $r(w)$ a priori. In addition, the choice of the cost function to minimize is somewhat arbitrary.

Samsung Exhibit 1022, Page 95 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Typically in connectionless networks, the network nodes will determine where to route a packet so that it gets forwarded to its destination. In such an environment, the source places the address of the destination in a **header field** of each packet. Upon receiving a packet, a network node examines the destination address, and uses this to index into a lookup table to determine the outgoing link that the packet should be forwarded on.

In contrast, *source routing* describes the situation where the route that a packet travels is determined by the source. In this case, the source explicitly specifies what route the packet should take within a header field of the packet. Typically, this specification consists of the list of node addresses along the path that the packet should travel. This eliminates the need for large lookup tables within each network node. For example, link state routing can be used, where the sources compute shortest paths.

47.4 Routing in Virtual Circuit Switched Networks

In connection-oriented networks, sources make connection requests to the network, transmit data to the destinations through the connection, and then terminate the connection. Since connection requests are distributed in time, a routing decision is typically made for each connection request as it is made. In other words, the sequence of connection requests that will be made is unknown, and a routing decision for a connection is not made until the connection request occurs.

Shortest path routing is still a viable approach in connection-oriented networks. The network nodes calculate shortest paths to all possible destinations. When a connection request is made, the connection request is forwarded along a shortest path to the destination, and the switching nodes along the path can determine if they have the resources (e.g., bandwidth, buffers) to support the connection request. If all network nodes determine that they can support the request, the network accepts the connection, link weights are updated to reflect the newly formed connection, and data for the connection is forwarded along the established path. If a network node along the shortest path is not able to support the connection request, the network would then reject the connection request.

It is possible for the network to attempt to route a connection request along another path if the first attempted path fails. However, this should be done carefully, since routing a connection on a path with many hops may use up bandwidth that could be better used to support other connection requests. **Trunk reservation** techniques may be used to advantage in this regard.

Alternatively, the sources might be responsible for specifying a desired path in a connection request, which would be a form of source routing. Obviously in this case, the sources would need knowledge of the network topology in order to make intelligent choices for the routes they request.

In connection-oriented networks, the source need not place the destination address into each packet. Rather, it could put a *virtual circuit identifier* (VCI) in each packet, which labels the connection that the packet belongs to. This can result in a savings of bandwidth if there are a large number of potential destinations, and a comparatively smaller number of connections passing through each link. A given connection can have a different VCI assigned on every link along the path assigned to it. When a connection is set up through a node, the node assigns a VCI to the connection to be used on the outgoing link, which may be different from the VCI used for the connection on the incoming link. The node then establishes a binding between the VCI used on the incoming link, the VCI assigned on the outgoing link, as well as the appropriate output link itself. Upon receiving a packet, the network node reads the VCI used on the incoming link and uses this as an index into a lookup table. In the lookup table, the node determines the appropriate outgoing link to forward the packet on, as well as the new VCI to use. This is sometimes referred to as VCI translation.

47.5 Hierarchical Routing

In a network with a very large number of nodes, lookup tables for routing can get very large. It is possible to reduce the size of the lookup tables, as well as the complexity of route computation, by

Samsung Exhibit 1022, Page 96 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

using a hierarchical routing strategy. One approach to a hierarchical routing strategy is as follows. A subset of the network nodes are identified as *backbone* nodes. A link connecting two backbones is called a backbone link. The set of backbone nodes and backbone links is called the backbone network. The nonbackbone nodes are partitioned into clusters of nodes called domains. Each backbone node belongs to a distinct domain, and is called the *parent* of all nodes in the domain. The nodes within a domain are typically geographically close to one another. Routes between pairs of nodes in the same domain are constrained not to go outside the domain. Routes between pairs of nodes in different domains are constrained to be a concatenation of a path completely within the source domain, a path completely within the backbone network, and a path completely within the destination domain. Routes are computed at two levels, the backbone level and the domain level. Route computations at the backbone level need only information about the backbone network and can proceed independently of route computations at the domain level. Conversely, route computations at the domain level need only information about the specific domain for which routes are computed. Because of this decoupling, the problem of route computation and packet forwarding can be considerably simplified in large networks.

With a hierarchical routing approach it is often convenient to have a hierarchical addressing strategy as well. For example, for the hierarchical routing strategy just described, a node address could be grouped into two parts, the domain address and the address of the node within the domain. Packet forwarding mechanisms at the backbone level need only look at the domain part of the node address and packet forwarding mechanisms at the domain level can ignore the domain part of the node address. Of course, it is possible to have hierarchical routing strategies with more than two levels as described.

47.6 Flow Control in Datagram Networks

Since sources in a connectionless network do not interact with the network prior to the transmission of data, it is possible that the sources may overload the network, causing buffers to overflow and packets to be lost. The sources may become aware that their packets have been lost, and as a result retransmit the lost packets. In turn, it is possible that the retransmission of packets will cause more buffers to overflow, causing more retransmissions, and so on; the network may be caught in a traffic jam where the network throughput decreases to an unacceptable level. It is also possible that buffers may overflow at the destinations if the sources send data too fast. The objective of flow control in connectionless networks is to throttle the sources, but only as necessary, so that rate of packet loss due to buffer overflow is at an acceptably low level.

A number of schemes have been proposed for flow control in connectionless networks. One technique that has been proposed is that network nodes that experience congestion (e.g., buffer overflow) should send the network nodes causing the congestion special control packets, which in effect request that the offending nodes decrease the rate at which packets are injected into the network. This has been implemented in the Internet, and the control packets are called *source quench* packets. A practical problem with this approach is the lack of accepted standards that specify under what conditions source quench packets are sent, and that specify exactly how network nodes should react to receiving a source quench packet.

We now focus our discussion on a common flow control technique used on connectionless networks, whereby the network carries traffic predominately from connection-oriented protocols, and *window flow control* is exercised within the connection-oriented protocols.

Window flow control is a technique where a source may have to wait for acknowledgements of packets previously sent before sending new packets. Specifically, a *window size* W is assigned to a source, and the source can have at most W unacknowledged packets outstanding. Typically, window flow control is exercised end-to-end, and the acknowledgements are sent back to the source by the destination as soon as it is ready to receive another packet. By delaying acknowledgements, the destination can slow down the source to prevent buffer overflow within the destination.

Samsung Exhibit 1022, Page 97 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Another motivation behind window flow control is that if the network becomes congested, packet delays will increase. This will slow down the delivery of acknowledgements back to the source, which will cause the source to decrease its rate of transmissions. The window size W is typically chosen large enough so that in the absence of congestion, the source will not be slowed down by waiting for acknowledgements to arrive.

Although it is possible to prevent buffer overflow by choosing the window sizes small enough and appropriately managing the creation of connections, this is typically not how datagram networks operate. In many cases, packet loss due to buffer overflow is possible, and window flow control is combined with an **automatic repeat request (ARQ) protocol** to manage the retransmission of lost packets. In particular, the source will wait only a limited time for an acknowledgement to arrive. After this amount of time passes without an acknowledgement arriving, the source will *timeout* and assume that the packet has been lost, causing one or more packets to be retransmitted. The amount of time that the source waits is called the *timeout value.*

As already mentioned, retransmissions should be carefully controlled to prevent the network throughput from becoming unacceptably low. We now provide a explanation of how this is commonly handled within the current Internet, specifically within the TCP protocol. The explanation is oversimplified, but should give a rough idea of how congestion control mechanisms have been built into the TCP protocol. This is the primary method of flow control within the current Internet.

The first idea [Jain, 1986] is that a timeout indicates congestion, and sources should wait for the congestion to clear and not add to it. In the TCP protocol [Jacobson, 1988], each timeout causes the timeout value to be doubled. This is called *exponential backoff,* and is similar to what happens on ethernet networks. In practice, so that the timeout values do not get unacceptably large, exponential backoff is stopped when the timeout value reaches a suitable threshold. After acknowledgements begin to arrive at the source again, the source can reset the timeout value appropriately.

In addition to exponential backoff, the TCP protocol dynamically adjusts the window size in response to congestion. In particular, a congestion window W' is defined. Initially, W' is set to the nominal window size W, which is appropriate for noncongestion conditions. After each timeout, W' is reduced by a factor of two, until $W' = 1$. This is called *multiplicative decrease.*

The window size used by the TCP protocol is at most W', and could be less. The idea is that after a period of congestion, a large number of sources could still retransmit a large number of packets and cause another period of congestion. To prevent this, each time a timeout occurs, the window size is reset to one. Hence, a source is then allowed to send one packet. If the source receives an acknowledgement of this packet, it increases the window size by one, and hence is then allowed to send two packets. In general, the source will increase the window size by one for each acknowledgement it receives, until the window size reaches the congestion window size W'. This is called *slow start* and is also used for new connections. After the window size reaches W', the window size is increased by one for every roundtrip delay, assuming no timeouts occur, until the original window size appropriate for noncongestion conditions is reached. This has been called *slower start.*

47.7 Flow Control in Virtual Circuit Switched Networks

In connection-oriented networks, flow control, in a broad sense, can be exercised at two different levels. At a coarse level, when a connection request is made, the network may assess whether it has the resources to support the requested connection. If the network decides that it cannot support the requested connection, the request is blocked (rejected). This is commonly called *admission control.* A connection request may carry information such as the bandwidth of the connection that is required, which the network can use to assess its ability to carry the connection. Some sources do not formally require any specified bandwidth from the network, but would like as high a bandwidth as possible. Such sources are not acutely sensitive to delay; as an example, a file transfer might be associated with such a source. These types of sources have been recently named *available bit rate* (ABR) sources. If all sources are ABR, it may be unnecessary for the network to employ admission control.

Samsung Exhibit 1022, Page 98 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

At a finer level, when a connection is made, flow control may be exercised at the packet level. In particular, the rate as well as the burstiness of sources may be controlled by the network. We focus our discussion on flow control entirely at the packet level. Proposals have also been made to exercise flow control at an intermediate level, on groups of packets called *bursts*, but we do not discuss that here.

We first discuss flow control for ABR sources. Given that the bandwidth allocated to an ABR source is variable, the issue of fairness arises. There are many possible ways to define fairness. One popular definition of fairness is *max–min fairness*, which is defined as follows. Each link in the network has a given maximum total bandwidth that is allocated to ABR sources using that link. Each ABR source has an associated path through the network that is used to deliver data from the source to the associated destination. An allocation of bandwidths to the ABR sources is called max–min fair if the minimum bandwidth that is allocated is as large as possible. Given that constraint, the next lowest bandwidth that is allocated is as large as possible, and so on.

A simple algorithm to compute the max–min fair allocation of bandwidths to the ABR sources is as follows. A link is said to be saturated if the sum of the bandwidths of all ABR sources using that link equals the total bandwidth available to ABR sources on that link. Each ABR source is initially assigned a bandwidth of zero. The bandwidth of all ABR sources is increased equally until a link becomes saturated. The bandwidth allocated to each ABR source that crosses a saturated link is then frozen. The bandwidth of all remaining ABR sources are then increased equally until another link becomes saturated, and the bandwidth of each ABR source that crosses a newly saturated link is then frozen. The algorithm repeats this process until the bandwidth of each ABR source is frozen.

One way to impose flow control on ABR sources that has been proposed is to use window flow control for each ABR connection at the hop level. That is, each ABR connection is subject to a window flow control algorithm for each hop along the path of the connection. Thus, each ABR connection is assigned a sequence of window sizes which specify the window size to be used at each hop along the path for the connection. Each node along the path of a connection reserves buffer space to hold a number of packets equal to the window size for the link incoming to that node. Packet loss is avoided by sending acknowledgements only after a packet has been forwarded to the next node along the path. This scheme results in backpressure that propagates toward the source in case of congestion. For example, if congestion occurs on a link, it inhibits acknowledgements from being sent back on the upstream link, causing the buffer space allocated to the connection at the node feeding the upstream link to fill. This will cause that node to delay sending acknowledgements to the node upstream to that node, and so forth. If congestion persists long enough, the source will eventually be inhibited from transmitting packets. It has been shown that if each node transmits buffered packets from ABR connections that pass through it on a round robin basis, that if window sizes are sufficiently large, and that if each source always has data to send, then the resulting bandwidths that the ABR sources receive are max–min fair [Hahne, 1991]. This scheme is also known as *credit-based flow control* [Kung and Morris, 1995], and algorithms have been proposed for adaptively changing the window sizes in response to demand.

Another way to achieve max–min fairness (or fairness with respect to any criterion for allocated bandwidths) is to explicitly compute the bandwidth allocations and enforce the allocations at the entry points of the network. Explicit enforcement of bandwidth allocations is known as *rate-based flow control*.

Bandwidth is defined in a time-average sense, and this needs to be specified precisely in order to enforce a bandwidth allocation. Suppose packet sizes are fixed. One way to precisely define conformance to a bandwidth allocation of ρ packets per unit time is as follows. In any interval of time of duration x, a source can transmit at most $\rho x + \sigma$ packets. The packet stream departing the source is then said to be (σ, ρ)*smooth*. The parameter σ is a constant positive integer and is a measure of the potential burstiness of the source. Over short intervals of time, the bandwidth of the source can be larger than ρ, but over sufficiently large (depending on the value σ) intervals of time, the bandwidth is essentially at most ρ. The source or network can insure conformance to this by using *leaky bucket flow control* [Turner, 1986], defined as follows.

Samsung Exhibit 1022, Page 99 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

A source is not allowed to transmit a packet unless it has a permit, and each transmitted packet by the source consumes one permit. As long as the number of permits is less than σ, the source receives new permits once every $(1/\rho)$ units of time. Thus, a source can accumulate up to σ permits. A packet that arrives when there are no permits can either be buffered by the source or discarded. Each ABR source could be assigned the parameters (σ, ρ), where ρ is the allocated bandwidth of the source, and σ represents the allocated burstiness of the source, in some sense.

Non-ABR sources require a certain minimum bandwidth from the network for satisfactory operation, and may also require that latency be bounded. Examples are multimedia sources, interactive database sources, distributed computing applications, and real-time applications. Leaky bucket flow control can also be applied to non-ABR sources. The parameter ρ specifies the maximum average bandwidth of the source, and the parameter σ measures the potential burstiness of the source, in some sense. If leaky bucket rate control is applied to all of the sources in the network, delay and buffering requirements can be controlled, and packet loss can be reduced or eliminated. Some recent analyses of this issue are presented in Cruz [1995, 1991a, 1991b]. We now briefly describe one important idea in these analyses.

A packet stream may become more bursty as it passes through network nodes. For example, suppose a packet stream arriving to a node is (σ, ρ) smooth, and that each packet has a delay in the node of at most d seconds. Consider the packet stream as it departs the node. Over any interval of length x, say, $[t, t + x]$, the packets that depart must have arrived to the switch in the interval $[t - d, t + x]$, since the delay is bounded by d. The number of these packets is bounded by $\rho(x + d) + \sigma = \rho x + (\sigma + \rho d)$, since the packet stream is (σ, ρ) smooth as it enters the node. Hence, the packet stream as it departs the node is $(\sigma + \rho d, \rho)$ smooth, and is potentially more bursty than the packet stream as it entered the node. Thus, as a packet stream passes through network nodes, it can get more bursty at each hop, and the buffering requirements will increase at each hop.

We note that buffer requirements at intermediate nodes can be reduced by using rate control (with buffering) to govern the transmission of packets *within* the network [Cruz, 1991b; Golestani, 1991]. This limits the burstiness of packet flow within the network, at the expense of some increase in average delay. It is also possible to reduce delay jitter by using this approach.

Defining Terms

ARQ protocol: An automatic repeat request (ARQ) protocol ensures reliable transport of packets between two points in a network. It involves labeling packets with sequence numbers and management of retransmission of lost or errored packets. It can be exercised on a link, or on an end-to-end basis.

Datagram network: A packet switched network that treats each packet as a separate entity. Each packet, or datagram, is forwarded to its destination independently. The data sources do not coordinate with the network or the destinations prior to sending a datagram. Datagrams may be lost in the network, without notification of the source.

Header field: A packet is generally divided into a header and a payload. The payload carries the data being transported, whereas the header contains control information for routing, sequence numbers for ARQ, flow control, error detection, etc. The header is generally divided into fields, and a header field contains information specific to a type of control function.

Packet switched network: A network where the fundamental unit of information is a packet. To achieve economy, the communication links of a packet switched network are statistically shared. Buffering is required at network nodes to resolve packet contentions for communication links.

Trunk reservation: A technique used in telephone networks whereby some portion of a communications link is reserved for circuits (trunks) that make most efficient use of the link. A link may reject a circuit request because of trunk reservation, so that other circuits which make better use of a the link may later be accepted.

Samsung Exhibit 1022, Page 100 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Virtual circuit switched network: A packet switched network that organizes data packets according to the virtual circuit that they belong to. Packets from a given virtual circuit are routed along a fixed path, and each node along the path may reserve resources for the virtual circuit. A virtual circuit must be established before data can be forwarded along it.

References

Bertsekas, D. and Gallager, R. 1992. *Data Networks*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ.

Comer, D. 1991. *Internetworking with TCP/IP*, Vol. I: *Principles, Protocols, and Architecture*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ.

Cruz, R.L. 1995. Quality of service guarantees in virtual circuit switched networks. *IEEE J. Selec. Areas Commun.*, 13(6):1048–1056.

Cruz, R.L. 1991a. A calculus for network delay, Part I: Network elements in isolation. *IEEE Trans. Inf. Th.*, 37(1):114–131.

Cruz, R.L. 1991b. A calculus for network delay, Part II: Network analysis. *IEEE Trans. Inf. Th.*, 37(1):132–141.

Cruz, R.L. and Tsai, J.T. 1996. COD: Alternative architectures for high speed packet switching. *IEEE/ACM Trans. Netwk.*, 4(1):11–21.

Golestani, S.J. 1991. Congestion-free communication in high speed packet networks. *IEEE Trans. Commun.*, 39(12):1802–1812.

Hahne, E. 1991. Round-robin scheduling for max-min fairness in data networks. *IEEE J. Selec. Areas Commun.*, 9(7):1024–1039.

Jacobson, V. 1988. Congestion avoidance and control. In *Proc. ACM SIGCOMM '88*, Stanford, CA, pp. 314–329.

Jain, R. 1986. A timeout-based congestion control scheme for window flow-controlled networks. *IEEE J. Selec. Areas Commun.*, 4(7):1162–1167.

Kung, H.T. and Morris, R. 1995. Credit-based flow control for ATM networks. *IEEE Netwk. Mag.*, 9(2):40–48.

Maxemchuk, N. 1987. Routing in the Manhattan street network. *IEEE Trans. Commun.*, 35(5):503–512.

Steenstrup, M., ed. 1995. *Routing in Communication Networks*. Prentice–Hall, Englewood Cliffs, NJ.

Turner, J.S. 1986. New directions in communications. *IEEE Commun. Mag.*, 24(10):8–15.

Further Information

Routing in Communication Networks, edited by Martha E. Steenstrup is a recent survey of the field of routing, including contributed chapters on routing in circuit switched networks, packet switched networks, high-speed networks, and mobile networks.

Internetworking with *TCP*, Vol. 1 by Douglas E. Comer provides an introduction to protocols used on the Internet, and provides pointers to information available on-line through the Internet.

Proceedings of the INFOCOM Conference are published annually by the Computer Society of the IEEE. These proceedings document some of the latest developments in communication networks.

The journal *IEEE/ACM Transactions on Networking* reports advances in the field of networks. For subscription information contact: IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway NJ, 088855-1331. Phone (800) 678-IEEE.

Samsung Exhibit 1022, Page 101 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

48

Transport Layer

48.1 Introduction 661
48.2 Transport Service 662
48.3 Data-Transfer Protocol 663
48.4 Connection-Management Protocol 665
48.5 Transport Protocols 667
48.6 Conclusions 670

A. Udaya Shankar
University of Maryland, College Park

48.1 Introduction

The transport layer of a computer network is situated above the network layer and provides reliable communication service between any two users on the network. The users of the **transport service** are e-mail, remote login, file transfer, web browsers, etc. Within the transport layer are **transport entities**, one for each user. The entities execute distributed algorithms, referred to as **transport protocols**, that makes use of the message-passing service of the network layer to offer the transport service. The most well-known transport protocols today are the Internet Transmission Control Protocol (TCP), the original transport protocol, and the ISO TP4.

Historically, transport protocol design has been driven by the need to operate correctly inspite of unreliable network service and failure-prone networks and hosts. The channels provided by the network layer between any two entities can lose, duplicate, and reorder messages in transit. Entities and channels can fail and recover. An entity failure is fail-stop, that is, a failed entity performs no actions and retains no state information except for stable storage. A channel failure means that the probability of message delivery becomes negligible; that is, even with retransmissions a message is not delivered within a specified time. These fault-tolerance considerations are still valid in today's internetworks.

To a first approximation, we can equate transport service to reliable data transfer between any two users. But reliable data transfer requires resources at the entities, such as buffers and processes for retransmitting data, reassembling data, etc. These resources typically cannot be maintained across failures. Furthermore, maintaining the resources continuously for every pair of users would be prohibitively inefficient, because only a very small fraction of user pairs in a network exchange data with any regularity, especially in a large network such as the Internet.

Therefore, a transport service has two parts: **connection management** and **data transfer**. Data transfer provides for the reliable exchange of data between connected users. Connection management provides for the establishment and termination of connections between users. Users can open and close connections to other users, and can accept or reject incoming connection requests. Resources are acquired when a user enters a connection, and released when the user leaves the

0-8493-8349-8/97/$0.00+$.50

© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 102 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

connection. An incoming connection request is rejected if the user has failed or its entity does not have adequate resources for new connections.

A key concern of transport protocols is to ensure that a connection is not infiltrated by old messages that may remain in the network from previous terminated connections. The standard techniques are to use the **3-way handshake** mechanism for connection management and the **sliding window** mechanism for data transfer within a connection. These mechanisms use **cyclic sequence numbers** to identify the connection attempts of a user and the data blocks within a connection. The protocol must ensure that received cyclic sequence numbers are correctly interpreted, and this invariably requires the network to enforce a **maximum message lifetime**.

In the following section, we describe the desired transport service. We then describe a simplified transport protocol that illustrates the core of TCP and ISO TP4. We conclude with a brief mention of performance issues and *minimum-latency* transport protocols.

Unless explicitly mentioned, we consider a **client–server** architecture. That is, the users of the transport layer are partitioned into clients and servers, and clients initiate connections to servers. Every user has a unique identity (id).

48.2 Transport Service

This section explains the correctness properties desired of the transport layer. We first define the notion of **incarnations**. An incarnation of a client is started whenever the client requests a connection to any server. An incarnation of a server is started whenever the server accepts a (potentially new) connection request from any client. Every incarnation is assigned an **incarnation number** when it starts; the incarnation is uniquely distinguished by its incarnation number and entity id.

Once an incarnation x of an entity c is started in an attempt to connect to an entity s, it has one of two possible futures. The first possibility is that at some point x becomes open and acquires an incarnation number y of some incarnation of s—we refer to this as "x becomes open to incarnation y of s"; at some later point x becomes closed. The second possibility is that x becomes closed without ever becoming open. This can happen to a client incarnation either because its connection request was rejected by the server or because of failure (in the server, the client, or the channels). It can happen to a server incarnation either because of failure or because it was started in response to a connection request that later turns out to be a duplicate request from some old, now closed, incarnation.

Because of failures, it is also possible that an incarnation x of c becomes open to incarnation y of s but y becomes closed without becoming open. This is referred to as a *half-open* connection.

A **connection** is an association between two open incarnations. Formally, a connection exists between incarnation x of entity c and incarnation y of entity s if y has become open to x and x has become open to y. The following properties are desired of connection management:

- Consistent connections: If an incarnation x of entity c becomes open to an incarnation y of entity s, then incarnation y is either open to x or will become open to x unless there are failures.
- Consistent data transfer: If an incarnation x of entity c becomes open to an incarnation y of entity s, then x accepts received data only if sent by y.
- Progress: If an incarnation x of a client requests a connection to a server, then a connection is established between x and an incarnation of the server within some specified time, provided the server does not reject x's request and neither client, server nor channels fail within that time.

Samsung Exhibit 1022, Page 103 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

- Terminating handshakes: An entity cannot stay indefinitely in a state (or set of states) where it is repeatedly sending messages expecting a response that never arrives. (Such *infinite chatter* is worse than deadlock because in addition to not making progress, the protocol is consuming precious network resources.)

Given a connection between incarnations x and y, the following properties are desired of the data transfer between x and y:

- In-sequence delivery: Data blocks are received at entity $y(x)$ in the same order as they were sent by $x(y)$.
- Progress: A data block sent by $x(y)$ is received at $y(x)$ within some specified time, provided the connection is not terminated (either intentionally or due to failures) within that time.

48.3 Data-Transfer Protocol

This section describes the sliding-window method for achieving reliable flow controlled data transfer, *assuming that users are always connected and correctly initialized.* Later we add connection management to this protocol.

Consider two entities c and s connected by unreliable network channels. The user at c produces data blocks to be delivered to the user at s. Because the channels can lose messages, every data block has to be retransmitted until it is acknowledged. For throughput reasons, entity c should be able to have several data blocks outstanding, that is, sent but not acknowledged. Similarly, entity s should be able to buffer data blocks received out-of-sequence (due to retransmissions or network reordering).

Let us number the data blocks produced by user c with successively increasing sequence numbers, starting from 0. The sliding window mechanism maintains two windows of sequence numbers, one at each entity.

Entity c maintains the following variables:

- na: $\{0, 1, \ldots\}$; initially 0. This is the number of data blocks sent and acknowledged.
- ns: $\{0, 1, \ldots\}$; initially 0. This is the number of data blocks produced by the local user.
- sw: $\{0, 1, \ldots, SW\}$; initially 0. This is the maximum number of outstanding data blocks that the entity can buffer.

Data blocks na to $ns-1$ are outstanding and must be buffered. The entity can accept an additional $sw - ns + na$ data blocks from its user, that is, data blocks ns to $na + sw - 1$. The sequence numbers na to $na + sw - 1$ constitute the **send window**, and sw is its size.

Entity s maintains the following variables:

- nr: $\{0, 1, \ldots\}$; initially 0. This is the number of data blocks delivered to the user.
- rw: $\{0, 1, \ldots, RW\}$; initially 0. This is the maximum number of data blocks that the entity can buffer.

Data blocks 0 to $nr - 1$ have been delivered to the user in sequence. A subset, perhaps empty, of the data blocks nr to $nr + rw - 1$ have been received and are buffered. The sequence numbers nr to $nr + rw - 1$ constitute the **receive window**, and rw is its size.

The easiest way for an entity to identify a received data block or acknowledgement is for the message to include the sequence number of the concerned data block. But such a sequence number field would grow without bound. Instead the protocol uses cyclic sequence numbers; that is, j mod N, for some N, instead of the *unbounded* sequence number j.

Samsung Exhibit 1022, Page 104 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

To ensure that a received cyclic sequence number is correctly interpreted, it is necessary for the network to enforce a maximum message lifetime, that is, no message older than the lifetime remains in the network. It then suffices if N satisfies

$$N \geq SW + RW + \frac{L}{\delta}$$

where SW and RW are the maximum sizes of the send and receive windows, L is the maximum message lifetime, and δ is the minimum time between successive data block productions.

Flow control is another issue in data transfer, that is, entity c should not send data faster than the network or entity s can handle. By dynamically varying the send window size, the sliding window mechanism can also achieve flow control. In particular, consumer-directed flow control works as follows: entity s dynamically varies its receive window size to reflect local resource constraints, and regularly requests entity c to set its send window size to the receive window size. Note that in this case, the previous condition on N reduces to $N \geq 2RW + L/\delta$.

We finish this section with a specification of the sliding window protocol. The data messages of the protocol have the form (D, *sid*, *rid*, *data*, cj), where *sid* is the sender's id, *rid* is the intended receiver's id, *data* is a data block, and cj is its cyclic sequence number.

The acknowledgment (ack) messages of the protocol have the form (ACK, *sid*, *rid*, cj, w), where *sid* and *rid* are as defined previously, cj is a cyclic sequence number, and w is a window size. When the message is sent, cj and w are set to the values of nr mod N and rw, respectively. Thus the message indicates the data block next expected in sequence. Because it acknowledges all earlier data blocks, it is referred to as a *cumulative ack*.

The events of the producer and consumer entities are shown in Fig. 48.1. There are two types of events. A *nonreceive* event has an enabling condition, denoted *ec*, and an action, denoted *ac*; the action can be executed whenever the event is enabled. A receive event for a message has only an action; it is executed whenever the message is received.

Entity c

```
Accept datablock from user (data)
 ec:  ns−na < sw
 ac:  buffer data as datablock ns;
      ns := ns+1

Send datablock (j)          // also resends
 ec:  0 ≤ j ≤ ns−na−1
 ac:  Send (D, c, s, datablock j, j mod N)

Reception of (ACK, s, c, cj, w)
 ac:  tmp := (cj−na) mod N;
      if 1 ≤ tmp ≤ ns−na then
        delete datablocks na to na+tmp
                    from send window;
        na := na+tmp;
        sw := w
      else if tmp = 0 then sw := max(sw, w)
      // else tmp > ns−na; do nothing
```

Entity s

```
Expand receive window
 ec:  rw < RW
 ac:  rw := rw+1

Deliver datablock to user (data)
 ec:  datablock nr in receive window
 ac:  data := datablock nr;
      nr := nr+1;
      rw := rw−1

Send acknowledgement
 ec:  true              // also resends
 ac:  Send (ACK, s, c, nr mod N, rw)

Reception of (D, c, s, data, cj)
 ac:  tmp := (cj−nr) mod N;
      if 0 ≤ tmp < rw then
        buffer data as datablock nr+tmp;
      // else tmp≥rw; do nothing
```

FIGURE 48.1 Events of sliding window protocol.

Samsung Exhibit 1022, Page 105 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

48.4 Connection-Management Protocol

This section describes a connection-management protocol. Traditional transport protocols, including TCP and TP4, identify successive incarnations by increasing, though not necessarily successively, incarnation numbers from some modulo-N space. Every entity uses a counter or a real-time clock to generate incarnation numbers for local incarnations.

Another feature of traditional transport protocols is that an entity stores a remote incarnation's number only while it is connected to the remote incarnation. This necessitates a 3-way handshake for connection establishment. A client that wants to connect to a server sends a connection request with its incarnation number, say, x. When the server receives this, it responds by sending a response containing x and a server incarnation number, say, y. When the client receives the response, it becomes open to y and responds by sending an ack containing x and y. The server becomes open when it receives the ack. The server could not become open when it received the connection request containing only x, because it may have been a duplicate from previous now terminated connection.

A 2-way handshake suffices for connection closing. An open entity sends a disconnect request that is acknowledged by the other entity.

A 2-way handshake also suffices for connection rejection. It is obvious that a server may have to reject a connection request of a client. What is not so obvious is that a client may have to reject a "connection request" of a server. Specifically, if a server receives an old connection request from a terminated incarnation of the client, the server attempts to complete the second stage of the 3-way handshake. In this case, the client has to reject the server.

The unreliable channels imply that a k-way handshake has the following structure: In every stage except the last, a message is sent repeatedly until the message of the next stage is received. The message of the last stage is sent only in response, otherwise the handshake would never terminate.

It is convenient to think of the protocol as a distributed system that is driven by user requests. Each user request causes the associated entity to initiate a 2- or 3-way handshake with the other entity. At each stage of the handshake, one entity learns something about the other entity and may issue an appropriate indication to its local user. At the end of the handshake, the protocol has served the user request. The protocol's behavior can be complex because two handshakes can be executing concurrently, with one of them conveying information that is relevant to the other.

There is an intricate relationship between the modulo-N space of the incarnation numbers and the handshaking algorithms, much more so than in the case of data transfer, since the latter assumes correctly initialized users. To achieve correct interpretation of received cyclic incarnation numbers, it is necessary to have bounds on message lifetime, incarnation lifetime, wait duration, and recovery duration. Under the reasonable assumption that the incarnation lifetime dominates the wait and recovery durations, it is sufficient and necessary to have

$$N \geq \frac{4L + I}{\alpha}$$

where L is the maximum message lifetime, I is the maximum incarnation lifetime, and α is the minimum time between successive incarnation creations at an entity. Most references in the literature incorrectly assume that $N \geq 2L/\alpha$ is sufficient.

This bound may not be satisfiable for exceedingly long-lived incarnations, say, of the order of days. In that case, if we assume that the probability of two successive connections having identical modulo-N client and server incarnation numbers is negligible (it is approximately $1/N^2$ under reasonable assumptions of incarnation lifetimes), then the following bound, which does not depend on I, suffices:

$$N \geq \frac{4L}{\alpha}$$

Samsung Exhibit 1022, Page 106 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

We now give a specification of the connection-management protocol.
A client entity maintains the following variables for each server s:

- *status*[s]: {CLOSED, OPENING, OPEN, CLOSING}; initially CLOSED. This is the status of the client's relationship with server s: CLOSED iff client has no incarnation involved with s. OPENING means client has an incarnation requesting a connection with s. OPEN means client has an incarnation open to s. CLOSING means client has an incarnation closing a connection with s.
- *lin*[s]: {NIL, 0, 1, ...}; initially NIL. This is the local incarnation number: NIL if *status*[s] equals CLOSED; otherwise identifies client incarnation involved with server s.
- *din*[s]: {NIL, 0, 1, ...}; initially NIL. This is the distant incarnation number: NIL if *status*[s] equals CLOSED or OPENING; otherwise identifies the incarnation of server s with which the client incarnation is involved.

A server entity maintains the following state variables for each client c:

- *status*[c]: {CLOSED, OPENING, OPEN}; initially CLOSED. This is status of server's relationship with client c: CLOSED iff server has no incarnation involved with c. OPENING means server has an incarnation accepting a connection request from c. OPEN means server has an incarnation open to c.
- *lin*[c]: {NIL, 0, 1, ...}; initially NIL. This is local incarnation number: NIL if *status*[c] equals CLOSED; otherwise identifies server incarnation involved with client c.
- *din*[c]: {NIL, 0, 1, ...}; initially NIL. This is the distant incarnation number. NIL if *status*[c] equals CLOSED; otherwise identifies the incarnation of client c with which the server incarnation is involved.

The messages of the protocol have the form (M, *sid*, *rid*, *sin*, *rin*), where M is the type of the message, *sid* is the sender's id, *rid* is the intended receiver's id, *sin* is the sender's incarnation number, and *rin* is the intended receiver's incarnation number. In some messages, *sin* or *rin* may be absent.

Each message is either a *primary* message or a *secondary* message. A primary message is sent repeatedly until a response is received or the maximum wait duration has elapsed. A secondary message is sent only in response to the reception of a primary message. Note that the response to a primary message may be another primary message, as in a 3-way handshake.

The messages sent by clients are as follows:

- (CR, *sid*, *rid*, *sin*): Connection request; sent when opening; primary message.
- (CRRACK, *sid*, *rid*, *sin*, *rin*): Acknowledgement to connection request reply (CRR); secondary message.
- (DR, *sid*, *rid*, *sin*, *rin*): Disconnect request; sent when closing; primary message.
- (REJ, *sid*, *rid*, *rin*): Reject response to a connection request reply that is received when closed. The *sin* of the received CRR is used as the value of *rin*; secondary message.

The messages sent by servers are as follows:

- (CRR, *sid*, *rid*, *sin*, *rin*): Reply to connection request in 3-way handshake; sent when opening; primary message.
- (DRACK, *sid*, *rid*, *sin*, *rin*): Response to disconnect request; secondary message.
- (REJ, *sid*, *rid*, *rin*): Reject response to a CR received when closed. The *sin* of the received message is used as the value of *rin*, secondary message.

The events of the client and server entities are shown in Figs. 48.2 and 48.3, assuming unbounded incarnation numbers.

Samsung Exhibit 1022, Page 107 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Client entity c: events concerning server s

```
ConnectRequest(s)
 ec: status[s] := CLOSED
 ac: status[s] := OPENING ; lin[s] := new incarnation number

DisconnectRequest(s)
 ec: status[s] = OPEN
 ac: status[s] := CLOSING

Abort(s)
 ec: status[s] ≠ CLOSED & "response timeout"
 ac: status[s] := CLOSED ; lin[s] := NIL ; din[s] := NIL

SendCR(s)
 ec: status[s] = OPENING
 ac: Send ( CR, c, s, lin[s] )

SendDR(s)
 ec: status[s] = CLOSING
 ac: Send ( DR, c, s, lin[s], din[s] )

Receive ( CRR, s, c, sin, rin )
 ac: if status[s] = OPENING & rin = lin[s] then
       status[s] := OPEN ; din[s] := sin ;
       Send ( CRRACK, c, s, lin[s], din[s] )
     else if status[s] = OPEN & rin = lin[s] & sin = din[s] then
       // duplicate CRR
       Send ( CRRACK, c, s, lin[s], din[s] )
     else if status[s] = OPEN & rin = lin[s] & sin>din[s] then
       // server crashed, recovered, responding to old CR
       Send ( REJ, c, s, sin ) ; status[s] := CLOSED ;
       din[s] := NIL ; lin[s] := NIL
     else if status[s] is CLOSED or CLOSING then Send ( REJ, c, s, sin )

Receive ( REJ, s, c, rin )
 ac: if status[s] is OPENING or CLOSING & rin = lin[s] then
       status[s] := CLOSED ; din[s] := NIL ; lin[s] := NIL
     // else status[s] is OPEN or CLOSED; do nothing

Receive ( DRACK, s, c, sin, rin )
 ac: if status[s] = CLOSING & rin = lin[s] & sin = din[s] then
       status[s] := CLOSED ; din[s] := NIL ; lin[s] := NIL
     // else status[s] is OPENING or OPEN or CLOSED; do nothing
```

FIGURE 48.2 Client events of connection management protocol.

48.5 Transport Protocols

A transport protocol between a client entity c and a server entity s consists of a connection-management protocol augmented with two data-transfer protocols, one for data transfer from c to s and another for data transfer from s to c. At each entity, the data-transfer protocol is initialized each time the entity becomes open and its events are executed only while open. The data-transfer messages are augmented by incarnation number fields, which are used by receiving entities to filter out data-transfer messages of old connections.

Samsung Exhibit 1022, Page 108 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Server entity s: events concerning client c

```
Abort(c)
 ec: status[c] ≠ CLOSED & "response timeout"
 ac: status[c] := CLOSED ; lin[c] := NIL ; din[c] := NIL

SendCRR(c)
 ec: status[s] = OPENING
 ac: Send ( CRR, s, c, lin[c], din[c] )

Receive ( CR, c, s, sin )
 ac: if status[c] = CLOSED & "rejecting connnections" then
        Send ( REJ, s, c, sin ) ;
     else if status[c] = CLOSED & "accepting connections" then
        lin[c] := new incarnation number ;
        status[c] := OPENING ; din[c] := sin
     else if status[c] = OPENING & sin>din[c] then
        // previous din[c] value was from some old CR
        din[c] := sin
     else if status[c] = OPEN & sin>din[c] then
        // client crashed, reconnecting
        if "willing to reopen" then
           lin[c] := new incarnation number ;
           din[c] := sin ; status[c] := OPENING
        else status[c] := CLOSED ; lin[c] := NIL ; din[c] := NIL
     // else status[c] = OPEN & sin ≤ din[c]; do nothing

Receive ( CRRACK, c, s, sin,, rin, )
 ac: if status[c] = OPENING & sin = din[c] & rin = lin[c] then
        status[c] := OPEN
     // else status[c] is OPEN or CLOSED; do nothing

Receive ( DR, c, s, sin, rin )
 ac: if status[c] = OPEN & sin = din[c] & rin = lin[c] then
        Send ( DRACK, s, c, lin[c], din[c] ) ;
        status[c] := CLOSED ; lin[c] := NIL ; din[c] := NIL
     else if status[c] = CLOSED then Send ( DRACK, s, c, rin, sin ) ;
     // else status[c] = OPENING; do nothing

Receive ( REJ, c, s, rin )
 ac: if status[c] = OPENING & rin = lin[c] then
        status[c] := CLOSED ; lin[c] := NIL ; din[c] := NIL
     // else status[c] is OPEN or CLOSED; do nothing
```

FIGURE 48.3 Server events of connection management protocol.

We illustrate with the protocols of the previous sections. Start with the connection-management protocol of Sec. 48.4 between c and s. Add two sliding window protocols, one from c to s and one from s to c, as follows:

- At each entity, introduce variables *ns, na, sw,* and the send window buffers for the outgoing data transfer, and *nr, rw,* and the receive window buffers for the incoming data transfer. These data-transfer variables are initialized whenever the entity becomes open. Whenever the entity becomes closed, these variables are deallocated.
- Modify the client as follows. Add *status*[*s*] = OPEN to the enabling condition of every data-transfer event. Add *sin* and *rin* fields to the sliding window protocol messages. When

Samsung Exhibit 1022, Page 109 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

a data-transfer message is sent, *sin* is set to the local incarnation number *lin*[*s*] and *rin* is set to the remote incarnation number *din*[*s*]. When a data-transfer message is received, first test for *status*[*s*] = OPEN, *sin* = *din*[*s*] and *rin* = *lin*[*s*]. If the test fails, ignore the message, otherwise, process the message as in the sliding window protocol specification.

- Modify the server similarly.

There are various ways to extend and integrate the data-transfer protocols.

The messages of the two data-transfer protocols can be combined. For example, the data messages sent by an entity can have additional fields to *piggy-back* acknowledgement information for incoming data, that is, fields for *nr* and *rw*. This is done in TCP and ISO TP4.

The preceding protocol uses cumulative acknowledgments. We can also use *reject* messages to indicate a gap in the received data. Rejects allow the data source to retransmit missing data sooner than cumulative acks. The protocol can also use *selective* acknowledgements to indicate correctly received out-of-sequence data. This allows the data source to retransmit only what is needed, rather than everything outstanding. Selective acks and rejects are not used in TCP and ISO TP4, although there are studies indicating that they can improve performance significantly.

The preceding protocol uses fixed-size data blocks. An alternative is to send variable-sized data blocks. This can be done by augmenting the data messages with a field indicating the size of the data block. TCP does this with an octet, or byte, as the basic unit. A similar modification would be needed for selective acks.

The connection-management protocol can be extended and integrated in several ways.

The preceding protocol allows either user to close a connection at any point, without waiting for data transfer to be completed. An alternative is so-called *graceful closing*, where a user can close only its outgoing data transfer. The user must continue to handle incoming data until the remote user issues a close also. TCP has graceful closing, whereas ISO TP4 does not. It is a simple matter to add graceful closing to a protocol that does not have it.

It is possible to merge connection establishment, data transfer, and connection termination. The connection request can contain data, which would be delivered after the server becomes open. The connection request can also indicate that after the data is delivered the connection is to be closed. TCP allows this.

TCP uses a single 32-b cyclic sequence number space to identify both incarnations and data blocks. When an incarnation is created at an entity, an **initial sequence number** is chosen and assigned to the incarnation. Successive new message sent by the incarnation, whether of connection management or data transfer, occupy increasing sequence numbers starting from this initial sequence number.

TCP messages integrate both data transfer and connection management. Every TCP message has fields indicating the sequence number of the message, the next sequence number expected, the data segment (if any), the segment length, and receive window size. A connection-management message that requires an acknowledgement is considered to use up a sequence number, say, n. This means that the next new connection-management message sent by the entity, whether or not it requires an acknowledgement, would have the sequence number $n + 1$. The remote entity can acknowledge a connection-management message by sending a message of any type with its next expected sequence number field equal to $n + 1$.

The TCP messages SYN, SYN-ACK, ACK, FIN, FIN-ACK and RST correspond, respectively, to the messages CR, CRR, CRRACK, DR, DRACK, and REJ of our protocol.

TCP provides **balanced opening**, a service that is outside the client–server paradigm. Here, if two entities request connections to each other at the same time, a *single* connection is established. In the client–server architecture, each entity would be considered to have a client-half and a server-half, and two connections would be established. In fact, TCP's algorithm for balanced opening is flawed: it may result in valid connection requests being rejected and invalid connection requests leading to connections. Fortunately, no application seems to use TCP's balanced-opening service. ISO TP4 does not have balanced opening.

Samsung Exhibit 1022, Page 110 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

48.6 Conclusions

This chapter has described the service expected of a transport layer and a distributed algorithm, or protocol, that achieves this service. The protocol faithfully illustrates the inner workings of TCP and ISO TP4. We conclude by outlining two currently active research areas in transport protocols.

One area is that of minimum-latency transport protocols. The delay in connection establishment incurred by the 3-way handshake is unacceptable for many transaction-oriented applications such as remote procedure calls (RPCs). Note that although transaction data can be sent with a connection request, the server cannot process the transaction until it confirms that this is a new request. This has motivated the development of transport protocols where the server can determine the *newness* of a connection request as soon as it is received, thereby achieving connection establishment with a 2-way handshake.

To achieve this, the server has to retain information about clients even when it is not connected to them. Consider a 3-way handshake between client incarnation x and server incarnation y. If the server had remembered the incarnation number, say, z, that the client had previously used when it connected to the server, then the server could determine that the connection request with x was new (because x would be greater than z). In that case, the server could have become open at once, resulting in a 2-way handshake connection establishment. A server cannot be expected to indefinitely remember the last incarnation number of every client to which it was connected, due to the enormous number of clients in a typical internetwork. However, a caching scheme is feasible, and several have been proposed, culminating recently in a proposed modification to TCP.

An alternative to caching is to use timer-based mechanisms. Here also, a server is required to maintain information on each client it has served for a duration comparable to that in cache-based mechanisms (the major component in both is the message lifetime). In most timer-based protocols, if a client's entry is removed before the specified duration, for example, due to a crash or memory limitation, then the server can incorrectly accept old connection requests of that client. Simple connection management protocol (SCMP) is an exception: by assuming synchronized clocks, it maintains correctness but it may reject new connections for a period of time depending on clock skews and other parameters. In any case, timer-based approaches do not have a backup 3-way handshake.

Another area of current research concerns the performance of transport protocols. As mentioned earlier, transport protocol design has been historically driven by the need to operate correctly inspite of unreliable network service and failure-prone networks and hosts. This has resulted in the premise that transport protocols should have minimal knowledge of the network state. For example, message roundtrip times are the only knowledge that TCP has of the current network state.

In recent years, Internet traffic has grown both in quantity and variety, resulting in increasing utilization, delays, and often congestion. This has prompted the development of retransmission and windowing policies to reduce network congestion, mechanisms to reduce processing time for transport protocol messages, etc. However, this performance work is still governed by the premise, motivated by fault tolerance, that a transport protocol should have minimal knowledge of the network state.

It is not clear whether this premise would be valid for the high-speed networks of the near and far future. Current implementations of TCP use a very conservative flow control scheme: the source reduces its send window by a multiplicative factor whenever roundtrip times indicate network congestion, and increases the send window by an additive factor in the absence of such indication. Although this is very robust, it tends to underutilize the network resources. It should be noted that most protocol proposals for high-speed networks make use of some form of resource reservation to offer varying quality-of-service guarantees. It is not clear how this can be incorporated within the current TCP framework, or even whether it should be.

Defining Terms

3-way handshake: A handshake between two entities involving three messages, say $m1$, $m2$, $m3$. The initiating entity repeatedly sends $m1$. The other entity responds by repeatedly sending $m2$. The initiating entity responds by sending a single $m3$ for each $m2$ received.

Samsung Exhibit 1022, Page 111 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Balanced-opening: Where two entities requesting connections to each other at the same time establish a single connection.

Client–server architecture: A paradigm where the users of the transport layer consist of clients and servers, connections are always between clients and servers, and connections are always initiated by clients.

Connection: An association between two open incarnations.

Connection management: That portion of the transport service and protocol responsible for the establishment and termination of connections between users.

Cyclic sequence numbers: The modulo-N sequence numbers used for identifying incarnations and data blocks.

Data transfer: That portion of the transport service and protocol responsible for the reliable exchange of data between connected users.

Incarnation: An instance of an entity created for a connection attempt.

Incarnation numbers: Sequence numbers for identifying incarnations.

Initial sequence numbers: The incarnation numbers of TCP.

Maximum message lifetime: The maximum duration for which a message can reside in the network.

Receive window: The data blocks that can be buffered at the consuming entity.

Send window: The data blocks that can be outstanding at the producing entity.

Sliding-window mechanism: A method for sending data reliably over unreliable channels, using cyclic sequence numbers to identify data, a send window at the data producer, and a receive window at the data consumer.

Transport entities: The processes within the transport layer.

Transport protocols: The distributed algorithm executed by the transport entities communicating over the network service.

Transport service: The service expected of the transport layer.

Further Information

The specification of TCP can be found in:

Postel, J. 1981. *Transmission Control Protocol: DARPA Internet Program Protocol Specification.* Request for Comment RFC-793, STD-007, Network Information Center, Sept.

The specification of ISO TP4 can be found in:

International Standards Organization 1989. *Information Processing Systems – Open Systems Interconnection – Transport Protocol Specification Addendum 2: Class four operation over connectionless network service.* International Standard ISO 8073/DAD 2, ISO, April.

The full specification and analysis of the data-transfer protocol in Sec. 48.3, including selective acks and rejects, can be found in:

Shankar, A.U. 1989. Verified data transfer protocols with variable flow control, *ACM Trans. Comput. Syst.*, 7(3):281–316.

The full specification and analysis of the connection management protocol in Sec. 48.4, as well as caching-based protocols and references to timer-based protocols can be found in:

Shankar, A.U. and Lee, D. 1995. Minimum latency transport protocols with modulo-N incarnation numbers. *IEEE/ACM Trans. Networking*, 3(3):255–268.

Composing data transfer and connection management is explained in:

Shankar, A.U. 1991. Modular design principles for protocols with an application to the transport layer, *Proc. IEEE*, 79(12):1687–1709.

Balanced opening and the flaws of TCP are described in:

Murphy, S.L. and Shankar, A.U. 1991. Connection management for the transport layer: Service specification and protocol verification. *IEEE Trans. Commun.*, 39(12):1762–1775.

Samsung Exhibit 1022, Page 112 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

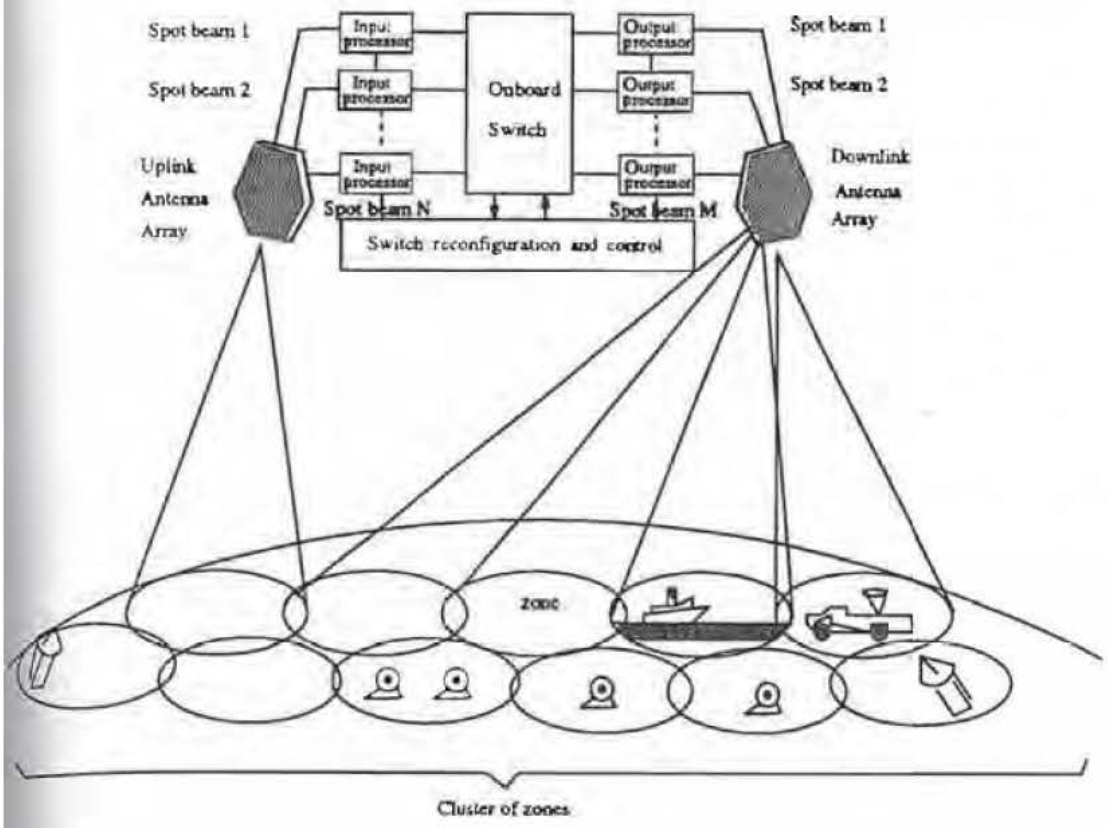


FIGURE 72.1 Onboard (OB) processing satellite using multiple-beam antennas and onboard switching.

2. Up- and downlinks are decoupled, thus allowing their separate optimization. For example, for stream (CS) and wideband Integrated Services Digital Network (WBISDN) users in a certain zone, time-division multiplexing (TDM) uplink and TDM downlink techniques are adequate. On the other hand, for a multiple of bursty Earth stations, in another zone, a **demand assignment multiple access–time division multiple access (DAMA–TDMA)** uplink/TDM downlink technique allows the optimal sharing of the uplink capacity. A third zone in heavy rain or under shadowing fading conditions might prefer a switching satellite that has code-division multiple access (CDMA) in the uplink and/or downlink.
3. The number of possible virtual connections is reduced from an order of N^2 to the order of N where N is the number of users.
4. Flexibility and reconfigurability includes, among other things, reprogramming of onboard processing (OBP) control memories, reconfiguration of Earth stations (time/frequency plans for each connection setup), accommodation of both circuit and packet switched traffic, adopting different flow and congestion control measures for each zone depending on current traffic loads and types, and accommodation of multirate and/or dynamically changing bit rate and and bit rate conversion on board.
5. SS has natural amenability to DAMA and ISDN services.
6. There is enhanced fault tolerance; a single satellite failure handling a cluster of zones can be easily replaced by another within hours or minutes thus averting the disasters of losing customers for a few days or months. A single failing antenna onboard the satellite can be easily replaced by another hopping antenna.
   These advantages come with a price, that is, the cost and complexity of the satellite itself.
7. SS increases the total capacity of the satellite system in terms of number of circuits and services provided within the same bandwidth.

Samsung Exhibit 1022, Page 113 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Applications

- Internetworking, low-cost and low-power user terminals, fixed site and mobile communications
- Video conferencing
- Interconnection of business oriented private networks, local area networks (LANs), metropolitan area networks (MANs), wide area networks (WANs)
- Narrowband and wideband ISDN
- Interconnection of all of the preceding to and through the public telephone network (PTN)
- Supercomputers networking
- Multiprotocol interconnection
- Universal personal communication (UPC)

72.2 Onboard Switching Types

- Reconfigurable static IF switching, also called **microwave switch matrix (MSM)** This has fixed (nonhopping) beams. The access techniques could be **frequency-division multiple access (FDMA)**, TDMA, or multifrequency-TDMA (MF-TDMA) giving rise to three kinds of switched satellites, that is, SS-FDMA, SS-TDMA, SS-MF-TDMA. The first, SS-FDMA, is the most common.
- Reconfigurable static intermediate frequency (IF) switching with regeneration and forward error correction but no baseband processing
- Baseband switching with fixed or hopping (scanning) spot beams
- Fast packet switching
- Photonic baseband switches
- Asynchronous transfer mode (ATM) oriented switches

Reconfigurable Static IF Switching

Both the uplinks and downlinks access techniques could be TDMA, FDMA, or MF-TDMA. However, the most commonly used are FDMA and TDMA on the uplinks and downlinks, respectively. This choice results in a reduced terminal ground station cost. FDMA is selected on the uplinks to eliminate the need for high-power amplifiers (HPA) at the user terminal [an FDMA user transmits all of the time but with a fraction of the bandwidth (BW)]. On the other hand, to utilize the full power of the OB (HPA), it should operate near saturation, thus increasing the downlinks signal and allowing very small aperture terminals (VSAT) operation. Simply a high-power high bit rate TDM operation in the downlinks is efficient. No demodulation or error correction takes place onboard. Fixed beams are used, and the number of such beams in the uplinks is typically less than that of the downlinks. Stations use frequency synthesizers to transmit the information using an RF carrier corresponding to destination. This form of FDMA is called single channel per carrier (SCPC) and is appropriate for low rate users (64 kb/s). This carrier is demand assigned to the user call via a signalling channel (which could be ground based or an out of band SCPC channel). Onboard, the various RF carriers pass through the low noise amplifier then become down converted to IF (Fig 72.2). Demultiplexing separates the various user signals that form the inputs to the static MSM. These are routed through the MSM to their destinations. The connection pattern of the switch (which input connected to which output) is configured by the master ground station or ground network control center (NCC) control channel. However, some of the connections may be permanent to allow broadcasting. Taking the example of the advanced communications technology satellite (ACTS) system of NASA [Cashman, 1992], the uplinks access technique is TDMA not FDMA, and the MSM is controlled by a digital controller and a dual storage bank of memories. The MSM switch is configured according to the 5-kb/s, telemetry, tracking, and command signal (TT&C)

Samsung Exhibit 1022, Page 114 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

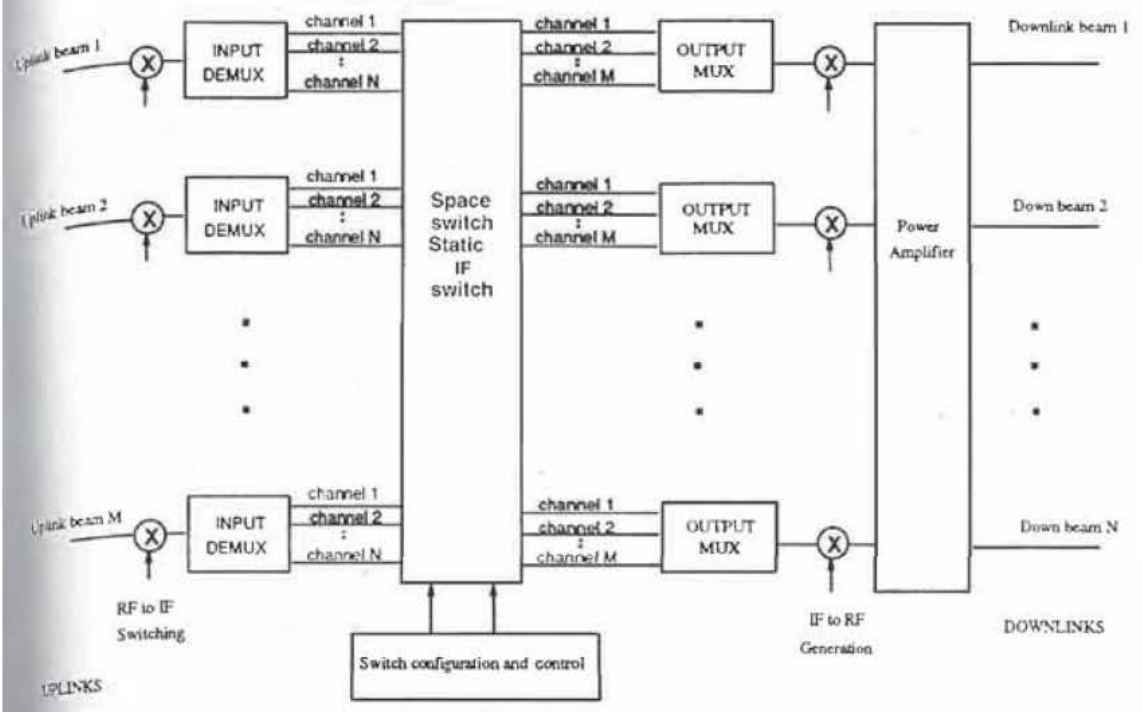


FIGURE 72.2 Nonregenerative onboard IF switching.

issued from the master ground station at the beginning of each superframe of 1-min duration. The TDMA frame is 1 or 32 ms and the switching time is approximately 100 ns. One bank of memory will be in active control of the switch while the other is waiting for TT&C new configurations. In the case of SCPC uplinks, the switch is configured on call basis (slower basis). Three active fixed or hopping beams exist on the MSM input and output. However, the MSM size is 4×4 to allow for redundancy. The output signals of the MSM switch going to same zone are grouped (multiplexed). The resultant is subsequently translated to RF frequency, amplified by solid-state power amplifiers (SSPA), and transmitted via output beams. An in-band or out-of-band signalling scheme is used by the user terminals to establish the call. The master control station finally assigns the right resources to the call, for example, one out of a pool on the FDMA uplinks and one time slot of a certain TDMA frame on the downlinks for that call. This reservation goes to the OBS, which becomes correspondingly configured for the call duration. We notice here the isolation between the up- and downlinks provided by OBS, for example, in the preceding case the uplink user transmits continuously using the assigned FDMA SCPC, i.e., disjoint from the TDM process taking place on the downlinks. Amplifying these FDMA signals OB creates intermodulation products (IMP). SCPC also suffers from its inability to meet variable bit rate (VBR) services. Multichannel per carrier (MCPC)/TDMA or MF/TDMA is typically employed to reduce these IMP. Here a number of users (terminals) are allocated channels (time slots) on a TDM frame, which is transmitted on a single carrier.

Other disadvantages of FDMA oriented MSM are: the difficulty of dynamic networks reconfiguration without data loss, the difficulty of operation with hopping beams, and broadcasting of multiple carrier transmissions to many destinations.

Regenerative IF Switching

This is similar to the previous static IF switching in the following respects: switching of IF signals, employing TDMA, FDMA, or MF-TDMA access techniques with the FDMA uplinks/MF-TDMA downlinks as the preferred technique, no baseband processing being involved, a station

Samsung Exhibit 1022, Page 115 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

simultaneously transmitting/receiving from many stations having to use a different carrier for each source (destination), programming reconfiguration via fixed connections and/or under the TT&C commands, use of low noise amplifiers (LNAs) and SSPAs, and RF to IF and IF to RF translations.

However, it differs in the following respects. The IF signals (following RF to IF translation and demultiplexing) are demodulated, forward error corrected, then forward error encoded [using error detecting and correcting coding (EDAC)] and then modulated [continuous phase shift keying (CPSK) or minimum shift keying (MSK)] again before feeding them to the MSM switch. If both uplinks and downlinks are of the MF-TDMA kind, then the various uplinks user TDM channels (slots) of each beam form the IF inputs to the switch. At the switch output, output IF slots are multiplexed again, and the resultant modulates the RF carrier of the downlinks beam. The regeneration involved in the demodulation–modulation processes allows the separation of uplinks and downlinks noises, fading, interference effects, and so forth. Also the decoding–encoding could be done at different rates thus providing for more protection against rain in the downlinks, for example, by decreasing forward error correction (FEC) encoding rate. These regenerations lead to better links performances and better amenability to MF-TDMA generation that needs higher transmission rates. Nevertheless, regenerative IF switching suffers the same aforementioned disadvantages of static IF switching.

Implementation of Regenerative and Static IF Switches via MSM Technology

Basically, this is a solid-state programmable crossbar switch. Signal switching at the matrix crosspoints is performed using dual-gate field-effect transistor (FET) active switches (employing GaAs technology for switches with IF bandwidths equal to or exceeding 2.5 GHz). In recent implementations, such as the NASA ACTS OBS [Cashman, 1992], input signal distribution and output signal combining are performed using passive recursive couplers (Fig. 72.3).

Among many alternatives, such as single pole, rearrangeable switching, and so forth, the coupler crossbar implementation was selected by both ACTS and other NASA systems for its simple control, low insertion loss, and amenability to broadcasting. In this architecture, a switching element [Fig. 72.3(b)] is situated between two directional couplers. If the data is to flow from input i to output j, the switch will be ON and the input and output couplers pass the data; otherwise, the switch is OFF and the data is reflected and absorbed by the coupler loads. To equalize the output power at all N input couplers connected to the N output lines, the coupler coefficients increase as the output number increases, that is, $\alpha_j = 1/(N - j + 2)$. The same logic α_j applies to the output coupler coefficient β_i as in Fig. 72.3(b). Switches at the various cross points are independently controlled to provide broadcasting as well as point-to-point connections. Moreover, the switch design could be modular both electrically and mechanically (as in NASA designs); for example, different switch points or parts of the switch can be repaired and tested independently, thus improving cost and reliability, or larger switches can be built from smaller modules.

The switch elements could be built using ferrite switches, p-i-n diodes, or single gate FETs. The first option is too slow, the second needs larger amounts of DC power, the third yields no gain during ON states; dual gate FETs are typically preferred. These use Schottky barrier diodes with very low junction capacitance, thus allowing the switching from gain to high insertion loss in about 100 ps. Two FETs are used, with the RF input signal applied to the first and the switch control applied to the second. This provides for excellent isolation between the DC control (which ranges from 30-dB loss to 10-dB gain) and the RF input. One could envision an IF switch with constant coupling coefficients (α_j, β_i) and variable switch amplifier gain. Typically, however, the earlier variable coupling gains implementation is preferred.

The switch states (configuration) and latency of each cross point are decided upon by ground control signals. This control unit is typically called control distribution unit (CDU). One bank of the dual control memory will be in active control of the switch states while the other bank receives the fresh ground control commands. These ground control signals also cause the exchange of the two memory bank identities to take account, for example, termination and/or new call establishments.

Samsung Exhibit 1022, Page 116 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

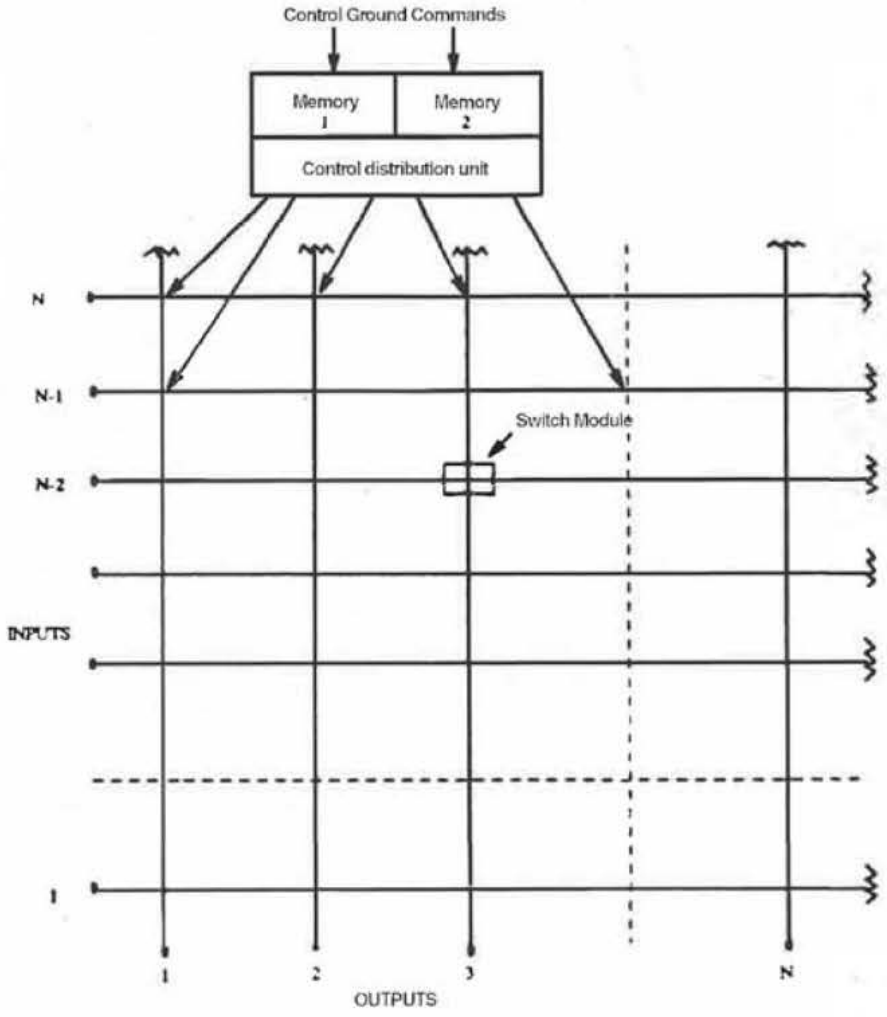


FIGURE 72.3(a) $N \times N$ IF switch matrix using crossbar configuration.

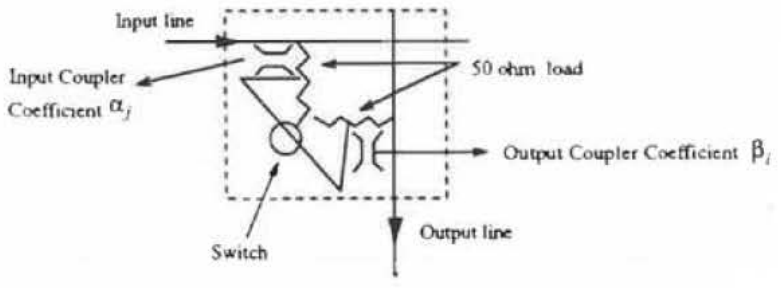


FIGURE 72.3(b) Switch module units: α_j = couplering coefficient of jth output cross point on the same input line; β_i = couplering coefficient of ith cross point on the same output line.

Samsung Exhibit 1022, Page 117 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The switch sequence (logical signals) that controls the switching state of all cross points is periodically dumped from the control memory till fresh control commands are loaded to the other memory bank from the ground control station, in which case the switch control signals are dumped from the other memory bank. The reconfiguration rate is a few microseconds and the switching time is in the 2–10 ns range.

The reliability of the IF switch matrix is highly dependent on the failure rate of the GaAs FET, which is roughly equal to 23×10^{-9} (including the associated passive elements). This reliability can be improved by adding extra rows and columns so as to provide alternate routes from inputs to outputs that do not pass by the failed cross points [the dashes of Fig. 72.3(a)]. Additional reliability could be achieved by using multiple ($N \times N$) switching planes in parallel at the expense of cost and weight. For the switch control unit, using the memory decoder yields a reliability of 0.8331 for 10 years operation. This can be improved further by using simple forward error correction techniques to correct for possible memory read or write errors.

For an $N \times N$ cross point IF switch and if the time frame has M data units, the control memory size is given by $M(N \log_2 N)$, since N cross points are controlled by $\log_2 N$ bits of memory (if decoding is used).

Implementation of Regenerative and Static IF Switches Using Optical Wave Guides

Future low orbit satellite clusters will probably communicate through intersatellite optical links at date rates of the order of few gigabits per second. This will make the instant and immediate switching of these signals in the optical domain a great asset compared to the classic tedious and slow approach of optical to electric conversion, switching of IF or baseband signals, then electrical to optical conversion at the switch outputs. In both IF and baseband implementations of optical switches, a directional coupler is the basic switching element.

The directional coupler (Fig. 72.4) consists of two waveguides situated close enough to allow the propagating modes (carrying the information signals to be switched) to interact (couple) and transfer power from one waveguide to another. Electrical control is still applied to allow the ON or OFF switching of the input signal to one or both outputs. This controls the index of refraction and, therefore, the coupling coefficients.

In the OFF state, a 3-dB optical isolation is achievable between outputs. If the switching voltage is 25 V, coupling lengths lie in the range of 1–2 cm, the loss associated with single mode fiber termination is of the order of a few decibels, and switching times lie in the 1–100 ns range. It is possible, however, to lower these to 50 ps by using traveling wave electrodes or by using free space optical switching techniques.

To reduce the N^2 complexity of MSM cross point switches using optical waveguides, different configurations are used. The one shown in Fig. 72.4(b) uses only $N(N-1)/2$, where $N = 6$ couplers, thus saving, for example, $(36 - 15) = 21$ couplers for a (6×6) switch.

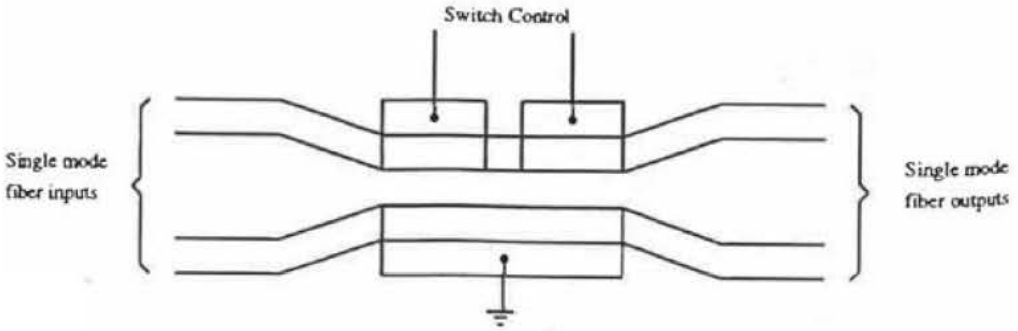


FIGURE 72.4(a) A 2×2 all optical switch module using directional coupler.

Samsung Exhibit 1022, Page 118 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

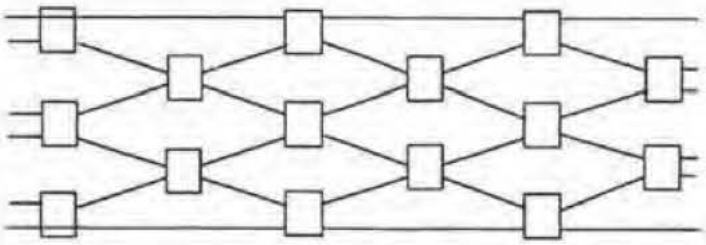


FIGURE 72.4(b) Multistage implementation of an optical $N \times N$ switch using $N(N-1)/2$.

Optical waveguides are not without disadvantages; for example, switch size is limited by the area of available substrate, high insertion loss (due to fan in losses of waveguides), high level of crosstalk, internal blocking in Benes, Banyan, or general multistage types, which eventually lead to difficulties in operation in broadcasting modes.

To enhance the reliability of optical wave guide IF switching architectures, replication in the case of $N \times N$ cross point structure or addition of a few input and output lines, and dilation (addition of more parallel lines between stages) in the case of Benes and multistage structures are generally recommended.

Onboard Baseband Switching

The uplink signals are downconverted, demodulated (possibly FEC decoded), switched, and then modulated, FEC encoded, upconverted, and transmitted downlink (Fig. 72.5). Switching takes place at baseband level, thus allowing for a much needed flexibility. For example, a station with only one carrier at hand can transmit to many destinations via different TDMA slots. These transmissions could reach the destinations on the different carriers, the different bit and FEC rates,

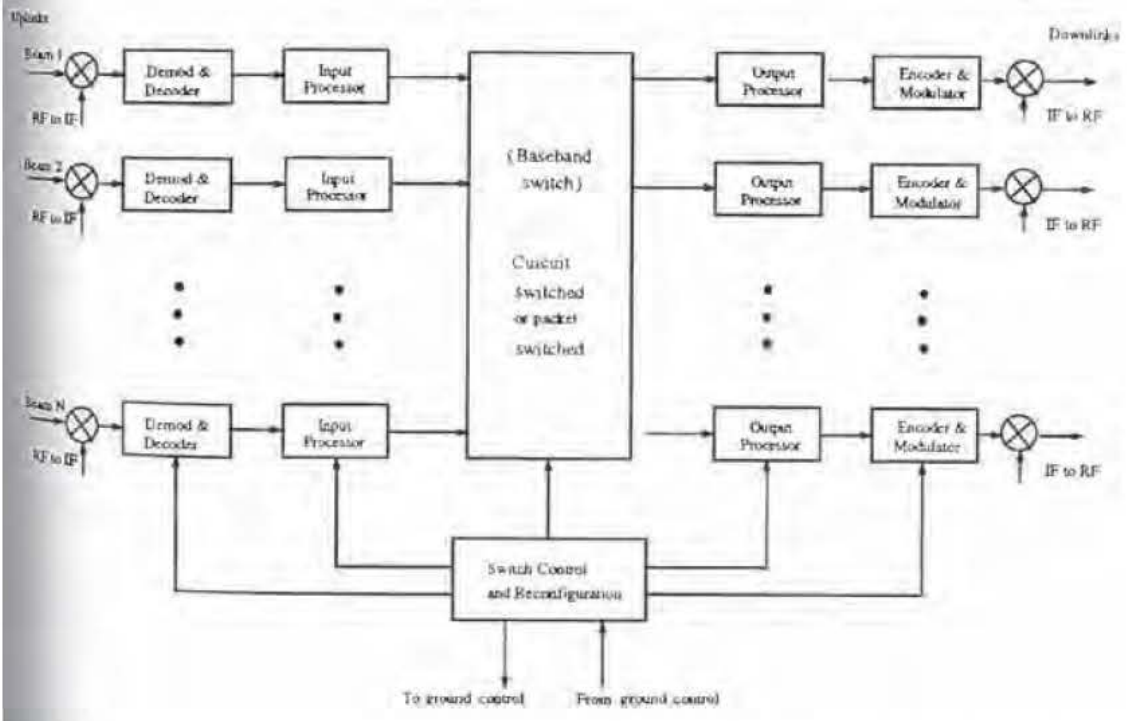


FIGURE 72.5 Building blocks of onboard baseband switching.

Samsung Exhibit 1022, Page 119 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

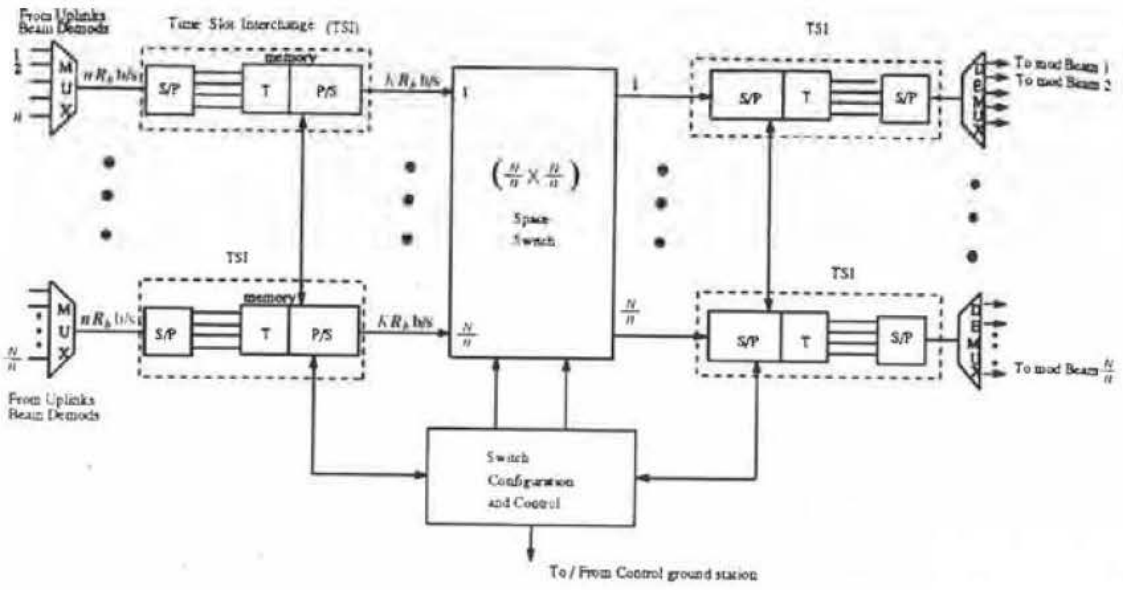


FIGURE 72.6(a) T–S–T baseband switching.

the different modulation formats, and so forth, that best suit the fading environment and other conditions in the downlink. Multicasting and accommodation of a mixture of high and low bit rates without needing multicarriers as in IF switching are some of the added benefits.

Baseband switching can be oriented toward circuit or packet oriented traffic. When large portions of the network traffic and/or most of the users are of the circuit switched type, and/or switch reconfiguration is infrequent and onboard, baseband circuit switching is more efficient. For a satellite network carrying mostly packet switched traffic, and/or circuit switched traffic with frequent configuration, and/or a combination thereof, a packet switch is more efficient. Hybrid packet/circuit baseband switching is also a viable alternative that has gained attention recently.

Baseband Circuit Switching Categories

Time–space–time (T–S–T), common memory (T-stage), distributed memory, and high-speed fiber optic ring are the main types of this category.

(T–S–T) Baseband Switching. This is used for multigigabit per second operation and has been adopted in the NASA ACTS system and for European Space Agency (ESA) prototypes. It has also been in use for some time in terrestrial networks (AT&T no. 4ESS switch). The basic idea is to reduce the space switch complexity by premultiplexing n of input data units (e.g., bytes), thus reducing the number of inputs and outputs of the space switch. However, this reduced switch has to switch data much faster, as will follow.

Following RF to IF conversion, filtering, and data demodulation and forward error decoding in each of the (N/n) uplink banks [Fig. 72.6(a)], the n baseband data units of each multiplexer (with a combined rate of nR_b b/s) are further multiplexed via the time slot interchange (TSI) box [Fig. 72.6(b)] to yield a bit rate at the TSI equal to kR_b b/s, where $k \geq 2n$ to allow for a strictly internally nonblocking switch. Internal nonblocking means that there are enough virtual paths inside the switch to connect any of the (N/n) inputs to any of the (N/n) outputs. The TSI is basically a serial to parallel converter followed by a T-stage (memory), which stores the data units of the uplink frames, followed by a parallel to serial converter. At the space switch output, another TSI executes a reversed sequence of events that finally feeds all data units of the downlink frame to the data forward error encoder and modulator, after which IF to RF conversion and subsequent transmission in the downlink takes place. To see the reduced complexity of this approach compared

Samsung Exhibit 1022, Page 120 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

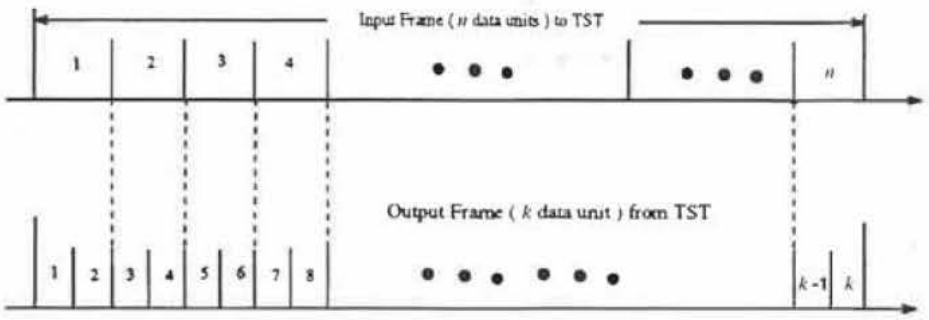


FIGURE 72.6(b) Time slot interchange (TSI) reading from T-stage is at least twice as fast as writing into this stage ($k \geq 2n - 1$).

to a conventional cross point switch, we take $N = 1024$, multiplexing of $n = 128$ slots (each can serve one unit of data), $k = 2n = 256$. The size of the space switch is $(N/n \times N/n)$, that is, (8×8) a great simplification compared to the corresponding conventional cross point switch of (1024×1024) complexity. The relation $k \geq 2n$ was derived in the 1950s for a class of three-stage Clos space switch, which resembles the T–S–T switch, and it effectively says that the speed of the space stage inputs (output of TSI unit) should be at least twice as fast as that of the TSI input [Fig. 72.6(b)]. Memory read and write from the T stages limit the speed of the external lines (from data demodulator and FEC decoders). If we take the frame time $T = 4$ ms, 1 byte as the data unit, input data rate of $R_b = 4$ Mb/s per input beam, the T-stage total memory size should be $2n \cdot T_F \cdot R_b \cong 4$ Mb. The total memory requirements for all input and output T-stages become $4(8 + 8) = 64$ Mb/s. The total switch throughput $= NR_b = 1024 \cdot 4$ Mb/s ≈ 4 Gb/s, which can be increased by increasing (N/n), that is, the number of pins of the space switch, and the memory access time per word of the input or output T-stages (assuming a 16-b-wide memory bus) is given by $1/(KR_b/16) = 16$ ns.

If current technology does not permit, one has to reduce n and/or the input data rate R_b. Dual memories (T-stages) of size $2n \cdot T_F \cdot R_b$ bytes are used on both input and output TSIs such that simultaneous read and write to the same memory can be achieved. Also, pipelined operation is assumed such that while bytes of the new frame are shifted at the input speed and stored in T-stages, bytes of the previous frames are shifted out to the output stage at the output speed.

To accommodate a multiplexity of user terminals of high and low speeds, additional multiplexing of lower rate users [not shown in Fig. 72.6(a)] is implemented at the input lines (prior to input TSI stage). This makes for a square $N/n \times N/n$ switch structure.

A sophisticated control processor (consisting mainly of memories and TSI timers) reconfigures the switch in a conflict free manner during the call establishment period for CS traffic and each frame for PS traffic. This consumes time and hardware since all three stages (input TSI, space switch, output TSI) have to be correspondingly adjusted. For example, if a user in input beam 3 has a call in slot 5 of the input frame to a user in output beam 6 who listens in slot 9 of the output frame, then the space switch control part makes sure that input beam (3) is connected to output beam (6) in each frame for one slot time. The TSI control at the switch inputs and outputs make sure that the input byte is written in the fifth location in input memory and the output byte read from the ninth location.

To improve T–S–T reliability, (N/n) active T-stages and r redundant ones are used in both input and output sides of the space switch, which is by itself replicated. Also error detection and correction can be used to protect the memories from permanent and transient errors. For example, 16 out of 20 (N/n out of $N/n + r$) redundancy and 1 out of 2 redundancy for space switch yields 99% reliability for 10 years.

It is also noted that the T–S–T architecture could in principle be used for routing packet switched traffic as well. However, switch control will be derived in this case from the individual input

Samsung Exhibit 1022, Page 121 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

packet headers, a process that slows the switch compared to the case of CS traffic. It is also generally assumed that all corresponding time slots of all inputs to TSIs are bit synchronized. This requirement will add costly hardware, especially in the case of hybrid PS/CS traffic; for example, some preambles will be stripped at the switch input, while some central headers will be routed through the switch. Furthermore, T–S–T switching has the advantage of expandability to higher speeds, and the disadvantage of encountering two switching TDM frame delays (corresponding to the input and output TSI stages).

Baseband Memory Switch (MS). This is the simplest and least costly architecture [Evans et al., 1986]. During a certain TDM frame, demodulated data bytes from all uplink beams are stored in a single double (dual) data memory (Fig. 72.7) in a sequential manner under the control of a set of counters or control memory. Parallel in time, the stored bytes from the previous switching TDM frame are served to the appropriate switch output under the supervision of the control memory. For one double dual memory to replace the many T-stages of the previous T–S–T architecture, this memory has to be large enough and fast enough to store the data bytes of the synchronized M input TDM frames in one frame time. A high-speed bus and a multiplexer (MUX) are thus used, and a demultiplexer (DEMUX) distributes the data memory contents into the various output channels under the control of simple counters and the configuration memory. Memory switches could be asymmetric (number of inputs $\neq$ number of outputs) and are internally nonblocking. The data memory size is $(2NS)$, where S is the number of TDM slots per frame and N is the number of output beams. This is half the memory requirements of the T-stages of T–S–T switches and is typically organized into 4 or 8 smaller memory blocks.

The contents of control memory serve as the addresses to the data memory and together with a simple counter associated with each output they read the right location of data memory in the right TDM slot of each output frame. The control memory is reconfigured with each new call. However, this is less intensive in time and in control hardware compared to the T–S–T case.

This architecture could possibly route packet switched traffic, in which case the switch control is derived from the destination in the packet header, which is more involved and time critical.

Bus speed and memory access time upperbounds the capacity of this architecture compared to the T–S–T that grows independently of these. Also, the memory is a reliability hazard avoided by

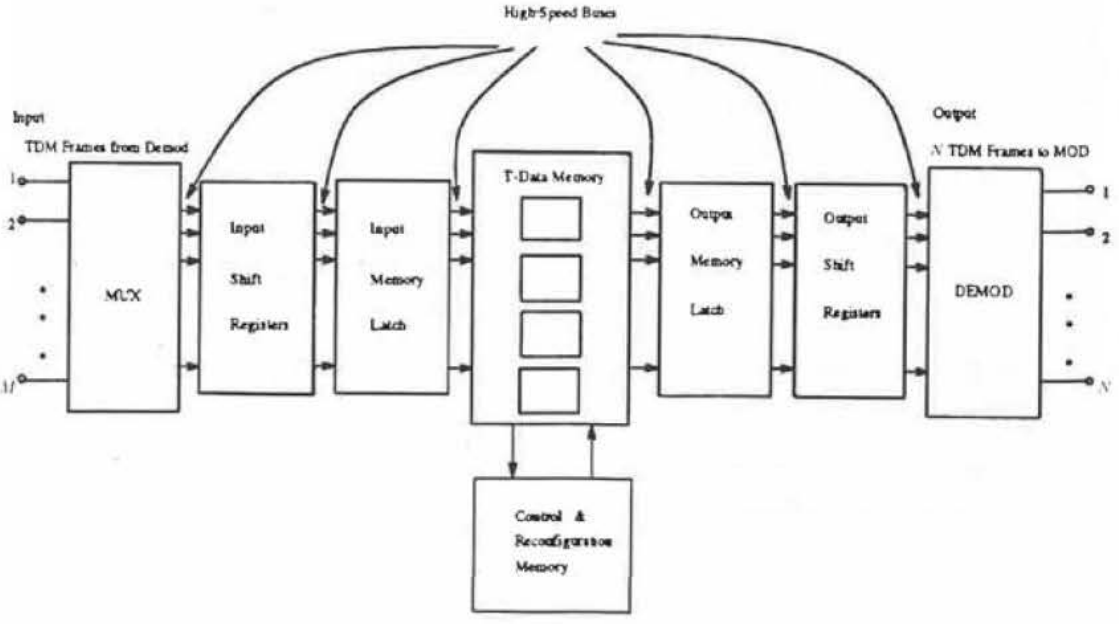


FIGURE 72.7 A baseband memory switch.

Samsung Exhibit 1022, Page 122 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

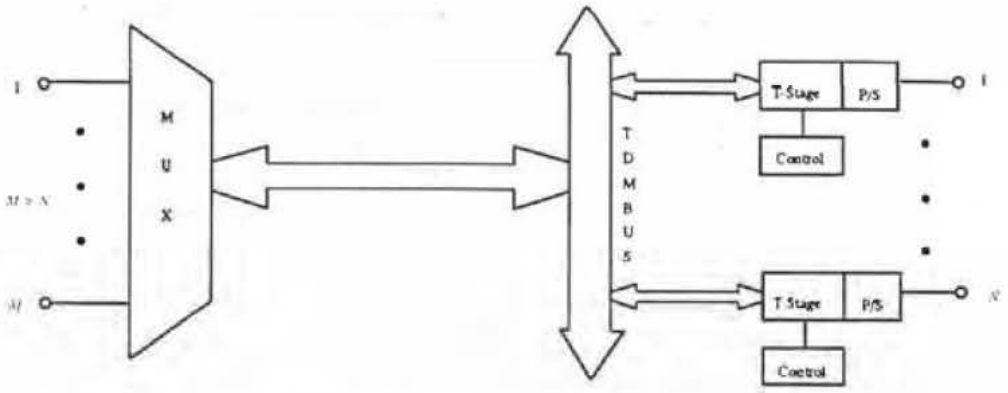


FIGURE 72.8 Distributed memory switch (DMS).

memory replication, and immunity to soft (intermittent) errors is provided by error detection and correction between data memory banks and read and write latches.

Power and weight are important considerations, and in recent designs (for example, the PSDE-SAT2 of the ESA), memory switches were approximately four times lighter than T–S–T switches and they took only 1/10th of the chips.

Distributed Memory Switch (DMS). To improve the reliability of the **memory switch (MS)** architecture, the memory could be split into N T-stages at the output side (Fig. 72.8). This way, failure of one or more of the T-stages means that only few (not all) beams are out of service.

Multiplexing of the uplink baseband TDM frames and sequential writing to the TDM bus takes place as in MS. All T-stages hear the broadcast TDM data; however, under distributed control and timers, only the appropriate data is sequentially written at the right time slot into the applicable T-stage. Reading data from T-stage to output beams is done via random access in different time slots of the output TDM frame. The *memory size* is still the same as in MS, and the distributed memory switch (DMS) is a strictly nonblocking switch. Redundancy is achieved by duplicating each T-stage; the MUX is also duplicated. Control units are also distributed and duplicated for fault tolerance. Reconfiguration of the T-stages is simple compared to T–S–T switches.

However, T–S–T architecture has better expansion capabilities than MS or DMS structures. Though bus speed is not a problem in the DMS, the bus is overloaded by the presence of the many T-stages, thus increasing the bus driver power consumption. Error detecting and correcting codes are also applied to the individual T-stages to protect against reading and writing errors.

Baseband Fast Packet Switches

Circuit switches are more or less amenable to handling packet switched traffic [Inukai and Pontano, 1993]. For example, the MS architecture becomes self-routing (and thus able to handle packet switched traffic) by stripping the headers of the packets that contain the destination and storing them in the control memory, which then selects the appropriate data unit from the data memory to the appropriate output register. In the case of the DMS architecture, these stripped packet headers are routed (using an address bus) to the distributed control units at the various output beams. A fiber optical token access scheme ring such as that of the fiber distributed data interface (FDDI) standard is directly adaptable to both CS and PS traffic.

MS, DMS, and T–S–T switch throughputs for PS traffic are reduced by the described header stripping, bus contention, memory size, and so forth. A self-routing multistage switch is generally preferred for handling PS traffic if throughputs higher than a few gigabits per second are required. A control memory is unnecessary in this case since part of the packet header is used to route the

Samsung Exhibit 1022, Page 123 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

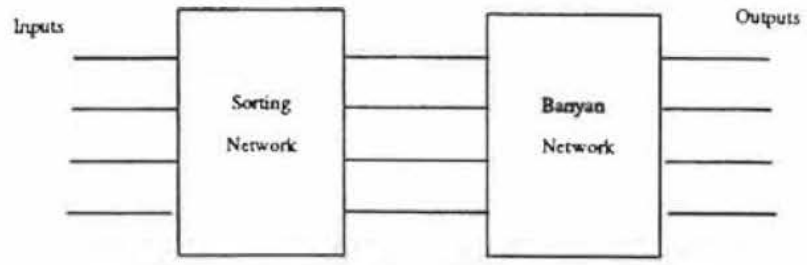


FIGURE 72.9(a) A sorter-Banyan baseband switch.

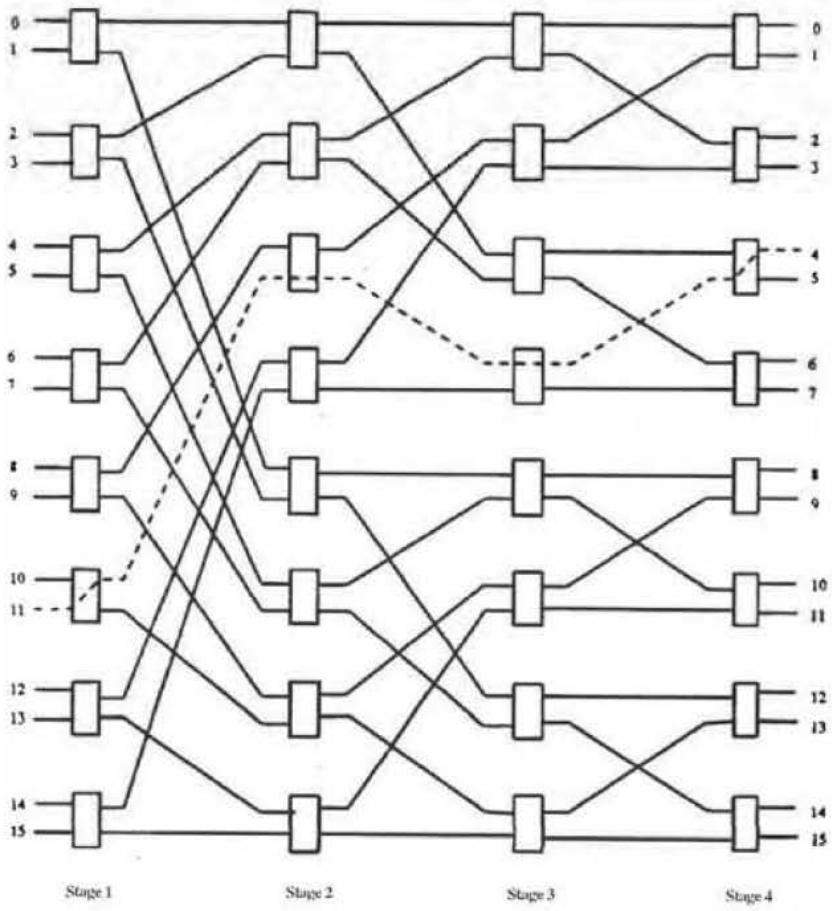


FIGURE 72.9(b) Banyan network with four stages and 16 input terminals. The dotted path corresponds to a packet header of 0100.

switch through each stage. Error detection and correction are used to protect against misrouting the packet due to header errors. However, complex data and control memories are not used; rather a set of cascaded (2×2) module switches is used. A typical multistage switch of the Banyan switch (BS) type is shown in Fig. 72.9(a) [the sorted **Banyan network** (**BN**) or Batcher Banyan network] [Hui and Arthurs, 1987].

The heart of this is the Banyan network (also called delta network), which is detailed in Fig. 72.9(b) for a 16×16 BN. The packet finds its path automatically by self-routing within the various stages in the $N \times N$ rectangular Banyan network, which is composed of ($\log_2 N = 4$) stages. Each stage

Samsung Exhibit 1022, Page 124 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

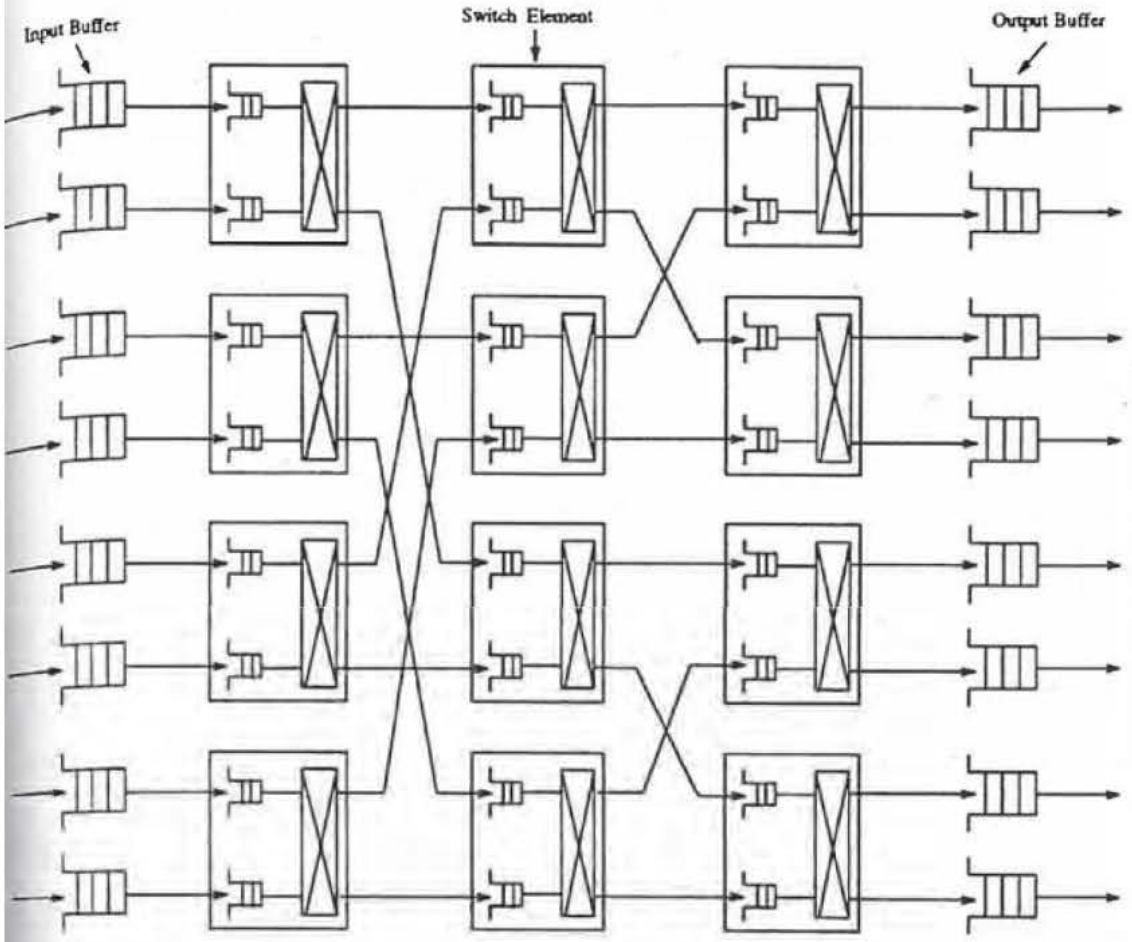


FIGURE 72.10 A buffered Banyan network.

consists of $(N/2)$ smaller switching elements. Each element is a small (2×2) binary switch. A control bit of 0 or 1 moves the input packet to the upper or lower element output, respectively.

A BN has the property that there is exactly one path from any input to any output but has the disadvantages of exhibiting internal blocking, that is, two packets from two inputs colliding at the same intermediate switching element output, even if they were destined to different final outputs. All of the previous CS described earlier have the advantage of being strictly nonblocking in this sense. However, both types (internally blocking and nonblocking switches) may have final output blocking, which is due to two packets from different inputs destined to same output. To solve the two problems of internal and output blocking, the sorting network of Fig. 72.9(a) should be used or one has to replace the whole figure by an alternate buffered Banyan network (Fig.72.10). In the latter, a limited buffering capacity exists at each (2×2) module input. If two input packets are destined to the same output, one packet is routed to the output and the other is temporarily stored for subsequent routing in the following time slot.

Among recent buffered Banyan implementations is the Caso–Banyan baseband switch of Toshiba [Shobtake and Kodama, 1990]. In that implementation, the FIFO property of the 2×2 switch element was violated such that when two head of line data units of the two input buffers contended for the same output, only one is forwarded to this output and another packet (not the head of line) of the other input is forwarded to the other output. It has been found that the unit data (packet, cell, byte) blocking rate is 1% of that given by conventional buffered Banyan networks at normalized input utilization factor $\rho = 0.85$. Now the sorting network approach of Fig. 72.9(a) [Hui and Arthurs, 1987] is used to reorder packets at the N input requests so they appear ordered according

Samsung Exhibit 1022, Page 125 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

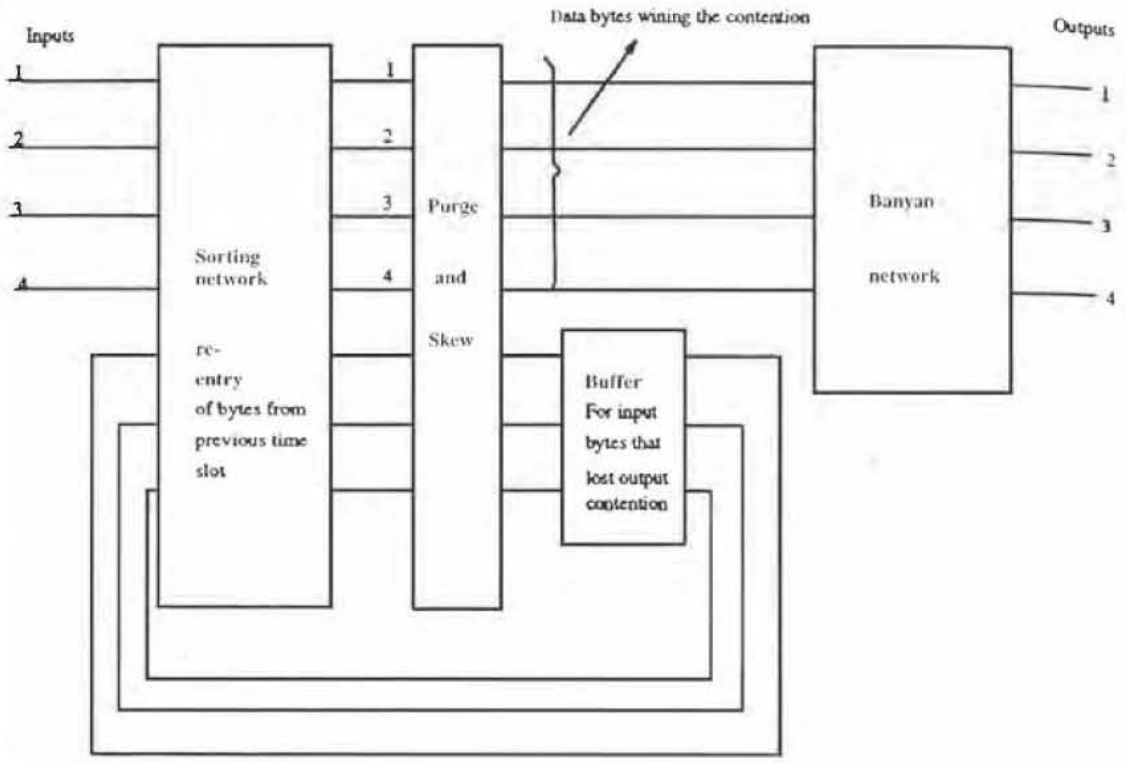


FIGURE 72.11 A Starlite network.

to their packet destination at the sorter N outputs (on the first phase). Also, in this phase only short connection requests are made by various inputs. This packet rearrangement guarantees no internal blocking when feeding these to the subsequent BN.

Output blocking is avoided by purging contending packets and acknowledging only the input requests that have priority (in the second phase), and forwarding only one of the sorter output packets to the BN (in the third phase).

Unacknowledged inputs will try retransmission of their output blocked packets in subsequent time slots. This three-phase algorithm slows the switching speed and lowers the achieved throughputs. Other implementations, such as the Starlite switch (Fig. 72.11), still have BN and use buffers at the output to store contending packets that did not win the output contention for subsequent routing in the following time slots, but do not have a separate request (phase 1) of the three-phase algorithm underlying Fig. 72.9(a). The Starlite approach [Huang and Knauer, 1984], however, has more hardware (buffers and control), and may produce packets out of sequence at the switch outputs.

Having realized that increasing the speed of Banyan type networks (reducing the blocking) was achieved only at the expense of increasing the hardware, a straightforward solution of *replicated-Banyan switches* provided modularity as well. Here, the number of internal paths (space diversity) is increased by distributing input traffic among many Banyan switches in different planes, thus reducing internal blockings. Output blocking may still be decreased by incorporating buffers at the various outputs of each switching plane. *Internal speed up* could be used to speed the data passage through the switch elements in all of the previous architectures (sorted Banyan, replicated Banyan, buffered Banyan, etc.) in a manner similar to that used in baseband CS (T–S–T, MS, DM). This leads to an almost strictly internal nonblocking switch as was previously outlined and a *space* switch component with less inputs and outputs.

Similar to baseband CS, the Banyan switches suffer from both *stuck at* and intermittent faults. To increase their reliability, replication of switching planes, *dilation* (addition of parallel links and more

Samsung Exhibit 1022, Page 126 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

switching stages within the same plane), and using EDAC codes to protect the packet headers before and after the baseband switches are the state-of-the-art techniques of fault tolerance in the baseband PS case [Kumar, 1989]. Also, self-testing hardware and testing sequences have been designed and added onboard to detect the presence of stuck at faults.

Gamma switching networks [sometimes called inverse augmented data manipulator (IADM)] have a multistage structure, which uses a 3×3 switching element and allows broadcasting, a property not readily available in the Banyan switching networks presented.

Omega and *close* switching networks also belong to the multistage group but could be made strictly nonblocking under certain conditions. Lack of amenability to broadcasting is the common thread in most of the Banyan-based and multistage networks. To have this capability, replicated or dilated Banyan networks are essential and/or *copy* networks could be used.

Photonic Baseband Switches

These are ultra fast and efficient switches that can handle a mix of CS and PS traffic and have an inherent broadcasting capability. TDM-based optical rings, optical token rings, **wave division multiplexing (WDM)** optical rings, and knockout switches are the basic types in this category. A single fiber-based ring can support a multitude of switched traffic to a total sum of a few gigabits per second. Optical technology (fibers, waveguides, free space optical transmission) provides a switching media whose bandwidths lie in the terahertz region, hence are bounded only by the processing speed of input and output interfaces, and stringent bit snchronization requirements around the ring, etc. Optical switch throughput exceeding 80 Gb/s can be accommodated with present day technology.

A TDM Optical Ring Switch

Each beam will have one input and one output processor on the optical ring see Fig. 72.12, which mainly shows the baseband stages). The input processor includes not only the optical transmitter,

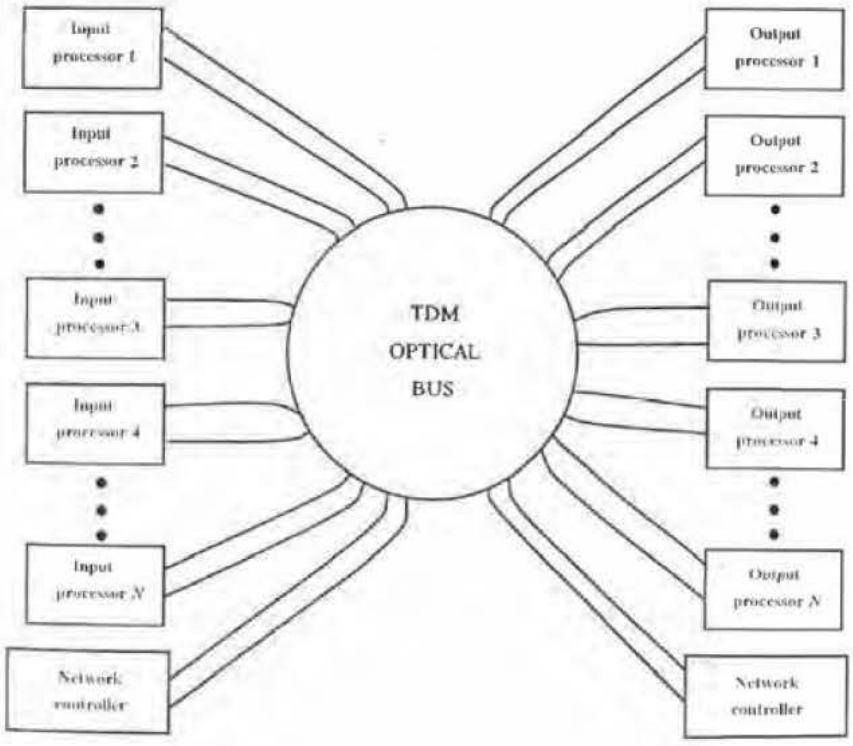


FIGURE 72.12 Onboard TDM optical ring switch.

Samsung Exhibit 1022, Page 127 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

but also EDAC decoding, packet assembly, descrambling, deinterleaving, burst and/or frame synchronization, input traffic demultiplexing, and possibly bit stuffing for CS traffic. The time frame on the optical ring is divided into slots (that serve, for example, one byte for CS traffic or one packet for PS traffic) and each input processor of an uplink beam is allocated a certain part of this ring switching frame (say, 8 slots) for its traffic to all output processors of the downlink. The optical ring traffic controller, which has a dedicated control packet at the ring TDM frame start, keeps the ring in sync, issues the frame start and ring management commands, and reconfigures (allocates slots) the TDM ring frame among the users (input processors).

For PS traffic the output processor of each beam receives all packets of the ring frame but processes only the ones destined to this beam by a filtering mechanism to recognize the destination address in each packet. For CS traffic, the output processor selects the appropriate slots of the frame based on the controller commands that are updated by the arrival of each CS call. Alternately, the input processor may insert dummy destination bits (addresses) in each CS packet only for routing purposes within the switch (TDM ring). These addresses will be filtered by the intended output processor. Note that dummy bits could be used even in the first approach to have the same packet size for both CS and PS traffic flowing through the ring so as to simplify synchronization and control hardware. Broadcasting is easily achieved by having all output processors identifying certain broadcast bits in the packet header and processing the packet accordingly.

The output processor of each beam will have the optical receiver, modulator, scrambler, interleaver, EDAC encoding circuits, and packet reassembly and multiplexing of the various packets on the TDM frame to the downlink. The reliability of the ring can be increased by using dual rings as in the FDDI standard, where failure of some parts or a whole ring does not affect the ring operation. Also, a bypass switch allows the isolation of faulty input and/or output processors. It is to be noted, however, that the access technique in FDDI is token ring and not TDM.

Wave Division Multiplexed (WDM) Baseband Switches

Figure 72.13 shows a WDM switch that uses a star coupler [Hinton, 1990]. This is a device with M optical inputs and N optical outputs that combines the M optical input signals, then divides them among the N optical outputs. Packetized baseband information is buffered at the various inputs. The input laser has a fixed wavelength. The output receiver should be tuned to the same wavelength as that of the intended input. Once both the transmitter and receiver lasers are triggered by the star controller, the input buffer serially shifts the data bits into the laser transmitter. This controller also

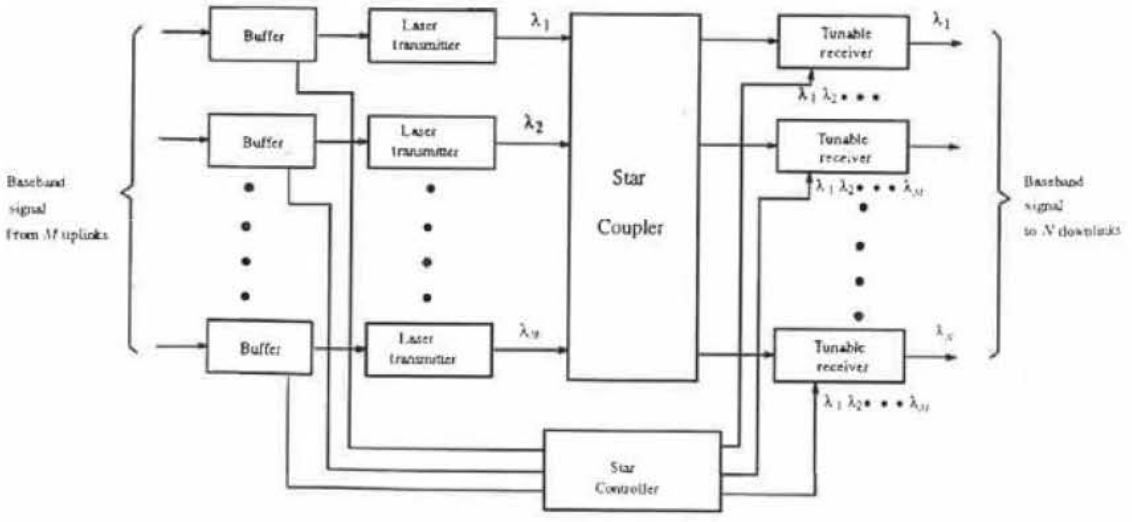


FIGURE 72.13 A WDM-based onboard baseband switch.

Samsung Exhibit 1022, Page 128 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

resolves the contention for switch outputs. This way, only one of the input packets is shifted, while the other contending input packet is left in the input buffer for the next time slot (packet time).

This represents the PS mode of operation and the amenability to broadcasting is clear, if a certain input packet wavelength is to be broadcast to a set of outputs, all involved output receivers are tuned to the same input wavelength. For CS operation the control commands are received from the ground station via out-of-band signalling channels. These reconfigure the star controller of the WDM switch by the arrival or termination of each new call. These commands may also be easily derived from a signalling TDM subframe slots (in-band signalling).

A mix of CS and PS traffic is easily accommodated, however, dummy bits would be inserted into CS traffic or the packet header stripped from PS traffic before entering the switch so as to have the same packet size passing through the WDM switch. Also, this switch becomes very efficient for directly switching CS intersatellite links in the optical domain.

For a cluster of low orbit satellites (LEO) (64 in the Global Star proposal [Louie, Rouffet, and Gilhausen, 1992]), optical intersatellite communications may render ground gateways (hubs) unnecessary, and will hence reduce the end-to-end user delay. However, with an efficient all optical switch such as WDM, we save the expensive and delay prone optical/electronic and electronic/optical conversions. Reconfiguration control of CS traffic is still done electronically by the arrival of each new call and does not have to be ultrafast. This could be achieved by an out-of-band signal emanating from the ground control station. This signal carries the time slot plan to each involved LEO, to enable it to communicate with its neighboring satellites. WDM technology has reached its maturity in terrestrial applications. Among WDM experiments is the HYPASS, where receiver wavelengths are fixed and the input laser transmitters are correspondingly tuned to the destination address. There is nothing in the weight, size, or cost that will forbid the use of these WDM switches in intersatellite switching applications so as to provide a total switch throughput over 50 Gb/s.

Asynchronous Transfer Mode-(ATM-) Oriented Switches

Onboard switches handling PS and CS traffic could become more involved with the passing of wideband ISDN and multimedia multirate services through the multibeam satellite. With the advent of ATM-based WBISDN, signals of different kinds (video, voice, data, facsimile, etc.) are all physically transported on a standard size packet called a *cell*.

Processing a standard size cell through the OB self-routing fast packet switch (multistage or Banyan types) makes the dummy bits addition and header stripping circuits mentioned earlier unnecessary. All services (PS, CS, variable bit rate, etc.) will have the same cell format. The cell has a header part and a data part. The cell size is 53 bytes, and the header mainly consists of two parts (Fig. 72.14). The first is the 12-b virtual path identifier (VPI) and the second is a 16-b virtual circuit identifier (VCI), which is followed by 12 b of error detection. The VPI is used to route the cell from satellite to next satellite or to ground, etc., enroute to the destination across the internetwork. The second cell field is a VCI between the two communicating entities. Cells find their way onboard the

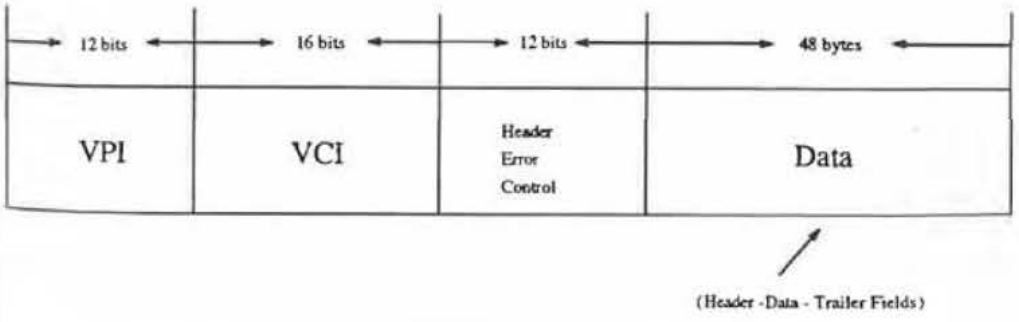


FIGURE 72.14 ATM cell user network interface (UNI).

Samsung Exhibit 1022, Page 129 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

satellite fast packet switch by a route identifier. This route identifier is obtained by feeding the VPI to a translation lookup table.

However, this translation process may become unnecessary if the VPI is configured such that its first few bits are the routing tag through the satellite packet type switch. This is one of many possible scenarios. If the VPIs are not permanently assigned to ground user stations, the satellite should also have OB translation tables (to obtain switching routing tag) for both the VPI and the VCI of each incoming cell, otherwise different cells from different source–destination pairs may have the same value in their VPI and VCI fields. This implies that the satellite will have the baseband routing switch (ABN type, for example) as well as an ATM switch. This complicates the OB processing, slows down the operation, and is not recommended. Thus far, we have assumed an ATM type of user station. For stations launching packets, not cells, to the satellite, the satellite should still have an ATM switch onboard and a packet processing facility to divide the packet into cells, to add VPIs and VCIs to cells, and so forth. For ground user stations launching CS traffic, the ATM switch onboard will assign a VPI, and a VCI to the incoming traffic and then route the traffic through the onboard baseband switch (BN, photonic, etc.). It is easily seen that unless the ATM users of the world have the ATM interface, the traffic arriving to the satellite during the early migration years of ATM technology may have cell, packet, or CS format thus complicating onboard processing. However, this might be inevitable if satellite vendors are to provide the same service to ground-based optical network users on standby or even on a competitive basis. Once ATM supersedes other switching modes OB processing will be less cumbersome except for the VPI and VCI processors referred to earlier.

72.3 Summary

Numerous switching techniques are in use or being developed for ground-based networks. We have tried to present the most relevant to the satellite environment. Knockout switches, token ring switches, and back to back hybrid packet/circuit switching, such as that of the ACTS system, are some of the techniques that could find application OB.

To compare the different switiching architectures, one has to look at their throughput blocking, queueing delay, and reliability performance (see the Further Information section).

Cost of hardware, weight, power consumption, and size of the OB switch are crucial considerations. Comparing the different architectures based on these could be a less quantifiable process. Comparing them on a basis such as amenability to migration to ATM environments, and to different uplinks and downlinks access techniques could be both art and science.

Defining Terms

Banyan network (BN): A self-routing switching technique for baseband signals.

Circuit switching (CS): A channel sharing technique where users are assigned physical or virtual circuits throughout their communications.

Demand assignment multiple access (DAMA): A multiaccess technique where each user is assigned a varying size time and/or frequency slot depending on his call needs.

Frequency-division multiple access (FDMA): A multiaccess techniques where each user is assigned a certain frequency band for his call.

Memory switch (MS): A baseband switching technique for OB processing.

Microwave switch matrix (MSM): A switching technique for microwave signals.

Onboard (OB): Part of the satellite payload.

Packet switching (PS): A transmission and networking technique where the data units are called packets.

Satellite switching (SS): A Satellite that radirects signals in the time, and/or frequency and/or space domains.

Samsung Exhibit 1022, Page 130 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Time-division multiple access (TDMA): A multiaccess technique where each user is assigned one or more slots in a certain time frame for his call.

Time space time (T-S-T): A baseband switching technique based on multiplexing in the time and space domains.

Wave-division multiplexing (WDM): A multiplexing technique for optical fiber networks also proposed for OB switching.

References

Cashman, W.F. 1992. ACTS multibeam communications package, description and performance characterization. In *Proc. AIAA 92 Conference*, AIAA, Washington, DC, pp. 1151–1161, March.

Evans, B.G. et al. 1986. Baseband switches and transmultiplexers for use in an on-board processing mobile/business satellite system. *IEE Proc.*, Pt. F, 133(4):336–363.

Hinton, S.H. 1990. Photonic switching fabrics. *IEEE Commun. Mag.*, 28(4):71–89.

Huang, A. and Knauer. 1984. Starlite—A wideband digital switch. In *Proc. IEEE GlobeCom-84*, pp. 5.3.1–5.3.5, Nov.

Hui, J.Y. and Arthurs, E. 1987. A broadband packet switch for integrated transport. *IEEE J. Selec. Areas Commun.*, SAC-5 (8):1264–1273.

Inukai, T., Faris F., and Shyy, D.J. 1992. On-board processing satellite network architectures for broadband ISDN. In *Proc. AIAA 92 Conference*, AIAA, Washington, DC, pp. 1471–1483, March.

Inukai, T. and Pontano, A.B. 1993. Satellite onboard baseband switching architectures. *European Trans. Telecomm.*, 4(1):53–61.

Karol, M.J., Hluchyj, M.G., and Morgan, S.P. 1987. Input versus output queueing on a space-division packet switch. *IEEE Trans. Com. on Comm.*, COM-35(Dec.):1347–1356.

Kumar, V.P. and Reibman, A.L. 1989. Failure dependent performance analysis of a fault-tolerant multistage interconnection network. *IEEE Trans. Comp.*, 38(12):1703–1713.

Kwan, R.K. et al. 1992. Technology requirements for mesh VSAT applications. In *Proc. AIAA 92 Conference*, AIAA, Washington, DC, pp. 1304–1314, March.

Louie, M., Rouffet, D., and Gilhausen, K.S. 1992. Multiple access techniques and spectrum utilization of the GlobalStar mobile satellite system. In *Proc. AIAA 92 Conference*, AIAA, Washington, DC, pp. 903–911, March.

Maronicchio, F. et al. 1988. Italsat—The first preoperational SS-TDMA system. In *Proc. IEEE Globe-Com 88 Conference*, pp. 53.6.1–53.6.5, March.

Oie, Y., Suda, S., Murata, M., and Miyahara, H. 1990. Survey of the performance of nonblocking switches with FIFO input buffers. In *Proc. IEEE ICC 90 Conference*, pp. 316.1.1–316.1.5, April.

Saadawi, T., Ammar, M., and Elhakeem, A.K. 1994. *Fundamentals of Telecommunications Networks*, Chap. 12, Wiley Interscience, New York.

Shabtake, Y. and Kodama, T. 1990. A cell switching algorithm for the buffered Banyan network. In *Proc. IEEE ICC 90 Conference*, pp. 316.4.1–316.4.7, April.

Further Information

Kwan et al. [1992] and Cashman [1992] provide a good treatment of the architectures and applications of microwave IF switches. Maronicchio et al. [1988], Evans et al. [1986] and Inukai and Pontano [1993] cover on-board baseband switches. The architectures and analysis of Banyan and other terrestrial baseband classes are investigated in Hui and Arthurs [1987], Huang and Knauer [1984], Carol, Hluchyj, and Morgan [1987], Oie et al. [1990], and Saadawi, Ammar, and Elhakeem [1994]. Photonic switching and LEO satellites are covered in Hinton [1990] and Louie, Rouffet, and Gilhausen [1992]. Fault tolerant switches are discussed in Kumar and Reibman [1989], which contains an exhaustive list of references.

Samsung Exhibit 1022, Page 131 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Mobile satellite systems share many aspects with other satellite systems as well as with terrestrial mobile radio systems (see Sec. 6.1). With respect to other satellite systems, the main differences are due to the small size and mobility of the user's terminal. The largest antennas used by mobile satellite terminals are 1-m parabolic dish antennas for large shipboard installations. At the other extreme are antennas small enough to be part of a handheld terminal. Also, the transmit power level of the terminal is limited due to small antenna size and often due to restrictions on available battery power. These factors have the consequence that mobile satellite systems operate at a significantly lower power margin than most mobile radio systems. Therefore, mobile satellite services are very limited unless there is line-of-sight transmission to the satellite.

74.2 The Radio Frequency Environment and Its Implications

When studying any radio system, it is generally instructive to begin by considering the basic link equation. This is particularly true for mobile satellite systems because the combination of limited electrical power on the satellite and at the mobile terminals and relatively small antennas on the mobile terminals results in a situation where transmitted power is a critical resource that should be conserved subject to the requirement of providing a sufficiently high signal-to-noise ratio to support the desired grade of service. Here, we only consider a simplified link equation with the objective of providing some context for subsequent issues. More detailed treatments of topics affecting link design are provided in Secs. 5.4 (satellite transmission impairments), 5.5 (satellite link design), and 5.6 (noise, antenna, and system temperature).

Consider the following simplified link equation:

$$C/N_o = P_t + G_t - L_p + G_r - T_r - k - L \text{ dB-Hz} \tag{74.1}$$

Where:

C/N_o = ratio of the signal power to the noise power spectral density after being received and amplified, dB-Hz
P_t = RF power delivered to the transmitting antenna, dBW
G_t = gain of the transmitting antenna relative to isotropic radiation, dBi
L_p = free-space path loss, dB
G_r = gain of the receiving antenna, dBi
T_r = composite noise temperature of the receiving system, dBK
k = Boltzmann's constant expressed in decibels, −228.6 dBW/K-Hz
L = composite margin accommodating losses due to a variety of transmission impairments, dB

A link equation of this type can be applied to either the **forward direction** or the **return direction**, as well as to the corresponding links between the satellite and the gateway or base station. Also, many satellites merely amplify and frequency translate the received signal from the uplink prior to transmission over the downlink. Therefore the end-to-end signal-to-noise ratio must incorporate both uplink and downlink noise. Further examination of some of the constituents of the link equation can provide some interesting insights. For instance, the transmitting antenna gain is a measure of an antenna's ability to convert the electrical radio signal into an electromagnetic waveform and then to focus that energy toward the receiving antenna. Similarly, the receiving antenna gain is a measure of an antenna's ability to collect the transmitted energy and then to convert the energy into an electrical signal that can be amplified and processed by the receiver. In general, the greater the gain of an antenna, the narrower is its beam. As an example, consider a circular parabolic antenna. Its antenna gain is given by the following expression:

$$G = 10 \log_{10}(\Omega \pi^2 D^2/\lambda^2) \text{ dBi} \tag{74.2}$$

Samsung Exhibit 1022, Page 132 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Where:

Ω = efficiency of the antenna
D = diameter of the antenna, m
λ = wavelength of the RF signal, m

Although the circular parabolic antenna is only one example of many possible antennas, this equation provides an order-of-magnitude indication of the size of antenna needed to provide a desired amount of gain.

The free space path loss is given by

$$L_p = 10 \log_{10}[(4\pi r)^2/\lambda^2] \text{ dB} \tag{74.3}$$

where r is the distance in meters between the terminal and the satellite.

Note that the path loss and both the transmitting and receiving antenna gains increase with the square of the radio frequency. From Eq. (74.1) this would imply that given constraints on the dimensions of the antennas, there is a net benefit in using relatively high radio frequencies. Although there is some truth to this, the benefits are not as great as one might think because the composite propagation losses also tend to increase with the radio frequency, as is described in Sec. 5.4 on satellite transmission impairments. Frequencies used for delivering mobile satellite services range from about 130 MHz up to 30 GHz! Use of the lowest frequency bands is dominated by low data rate store-and-forward services provided with low Earth-orbit (LEO) satellites. Most mobile satellite telephone services, with typical transmission rates ranging from several kilobit per second (kbps) up to 64 kbps, are provided in frequency bands located between 1500 and 2500 MHz. Interactive packet data and circuit-switched services, at transmission rates up to several megabit per second, are planned for the higher frequency bands. This correlation between transmission rates and carrier frequencies is a result of both the preceding link considerations as well as the fact that larger spectrum allocations are available in the higher frequency bands than in the lower ones.

Efficient use of the satellite's two critical resources, spectrum and power, can be achieved if the satellite's antenna system covers the appropriate area of the Earth's surface with multiple beams instead of one large beam. The total allocated system bandwidth is divided into a number of distinct subbands, which need not be of equal bandwidths, and each beam is then assigned a subband in such a way that some desired minimum isolation between beams with the same subband is maintained. This situation is illustrated in Fig. 74.1. Although not necessarily the case in an operational system, this example illustrates the beams as being circular and equal in area. Here, there are 4 distinct subbands and 16 beams. Note that in this example each frequency subband is used four times. In general, the frequency reuse factor is defined to be the ratio of the number of beams to the number of distinct frequency subbands. Improved power efficiency is a direct result of the higher satellite antenna gain implicit in the smaller beam area. Ideally, the satellite should be able to apportion dynamically its total transmit power among the individual beams depending upon the loading in each beam.

The satellite mobile channel is subject to multipath fading and shadowing just as the terrestrial mobile channel. The physical mechanisms for these propagation phenomena are the same. Shadowing is due to foliage, buildings, or terrain blocking the line-of-sight signal from the satellite. Many propagation campaigns have been performed to investigate the statistics of shadowing in mobile satellite environments. For example, Fig. 74.2 shows some results of data gathered near Ottawa, Canada, in June of 1983 [Butterworth, 1984]. Note that these curves exhibit a *knee* between 50 and 90%. For the portion of the curves to the left of the knee, the characteristics are dominated by multipath fading, whereas shadowing is the dominating factor to the right of the knee. With adequate margins, it is reasonable to expect mobile satellite systems to maintain communications links in environments characterized by fading and light shadowing, but it is not practical to overcome heavy shadowing for most services. For this reason, availability targets between 90 and 95%, corresponding to margins of less than 10 dB, are common for land mobile satellite systems

Samsung Exhibit 1022, Page 133 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

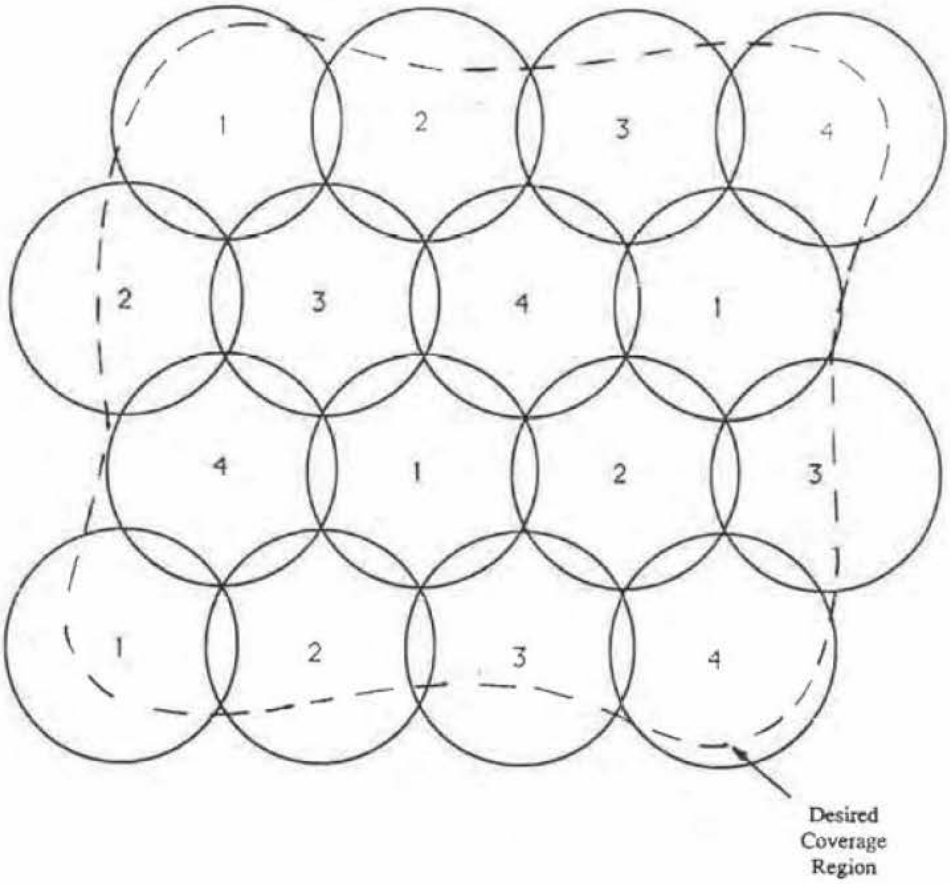


FIGURE 74.1 An example illustrating frequency reuse with a multiple-beam satellite. Here, the total allocated system frequency band is divided into four subbands. The assignment of the subbands to the beams is denoted by the number shown in each beam.

delivering services to users in rural, highway, and suburban environments. Availability figures for urban environments tend to be much lower. Exceptions include very low-capacity services, such as paging, which may require margins as high as 30 dB to allow for building penetration. Of course, the margin that is required to meet a given level of availability in a particular set of user environments is dependent on many factors including the elevation angle to the satellite, the type of terrain, the density and height of the vegetation, and the number of man-made obstacles.

The shadowing is often modeled as a log-normal process [Butterworth, 1984; Loo, 1984] where the direct path power (relative to the median power) has a distribution given by

$$P[S < x_{\mathrm{dB}}] = 0.5 + 0.5\,\mathrm{erf}\left(\frac{x_{\mathrm{dB}} - m_s}{\sqrt{2}\sigma_s}\right) \tag{74.4}$$

where S is the received signal strength in decibel, m_s is the average decibel shadowing loss, and σ_s is the decibel variance of the shadowing loss. The parameters of this distribution depend not only on environment but on the elevation angle to the satellite and on frequency. For a variety of situations, estimates of these parameters can be derived from the empirical results, of propagation campaigns, provided in ITU recommendations [1992, 1995] and in Stutzman [1988].

Multipath fading results from interference caused by reflections of the radio signal from the immediate environment, with the time-varying nature being caused by motion of the terminal

Samsung Exhibit 1022, Page 134 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

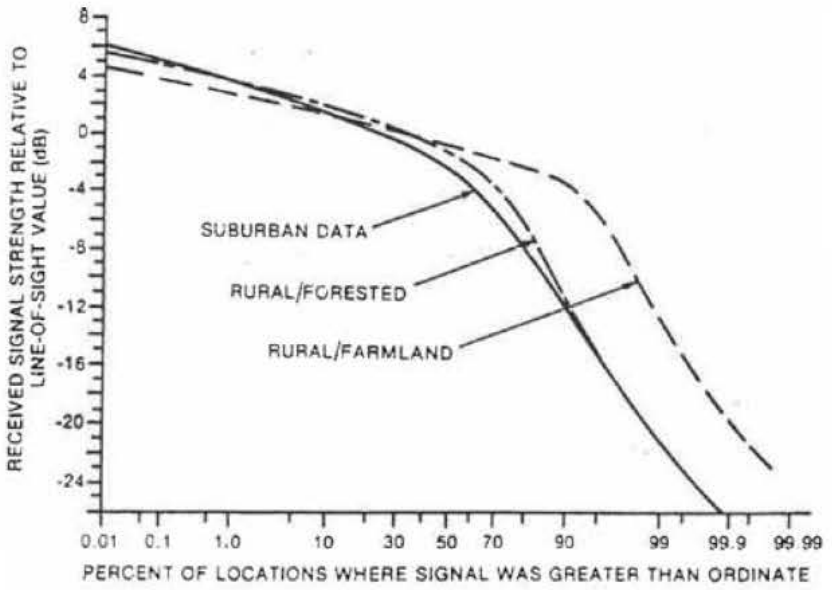


FIGURE 74.2 Distribution functions for data recorded near Ottawa, Canada, in June 1983. The measurements were made with an omnidirectional antenna at a frequency of 1542 MHz. The elevation angle (the angle between a plane parallel to the Earth's surface and the line-of-sight path to the satellite) was approximately 20°. (*Source:* Butterworth, 1984.)

through the environment. In satellite applications, the reflections are generally too weak to be useful by themselves but not so weak as to be ignored. Consequently, the satellite mobile channel is usually modeled as a Rician fading channel consisting of an unshadowed direct path and a diffuse component, which has a complex Gaussian distribution (Rayleigh distributed amplitude and uniformly distributed phase). The amplitude distribution of the resulting channel is the noncentral chi-squared distribution with two degrees of freedom (assuming a unit amplitude direct path)

$$P[R > r] = 1 - e^{-(a^2+b^2)/2} \sum_{k=0}^{\infty} \left(\frac{a}{b}\right)^k I_k(ab) \tag{74.5}$$

where $a = \sqrt{K}$, $b = r/\sigma$, and $I_k(\cdot)$ is the kth-order modified Bessel function. The quantity K is the ratio of the direct path power to the average power of the multipath $2\sigma^2$. Rician channels are characterized by their K factor. The probability of a fade of a given depth is shown in Fig. 74.3 for various K factors. Note that there is a greater than 50% probability that the diffuse component will increase the signal strength. K factors with vehicle-mounted antennas (automobile, aircraft, or ship) are typically 10 dB and higher, with the worst-case value of 10 dB occurring at low elevation angles. For handheld terminals, where the antenna has little directivity, lower K factors can be expected.

The time-varying nature of the fading, in particular, the duration of the fades, is characterized by its spectrum. For land–mobile applications the Clarke model, which assumes uniformly distributed scatterers in a plane about the moving terminal, is often used. The maximum Doppler frequency observed is related to the transmission frequency by the expression $f_d = v/\lambda$, where v is the velocity of the terminal, and λ is the transmission wavelength. The Clarke model for the diffuse spectrum is given by

$$S(f) = \frac{1}{\sqrt{1 - (f/f_d)^2}} \tag{74.6}$$

Samsung Exhibit 1022, Page 135 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

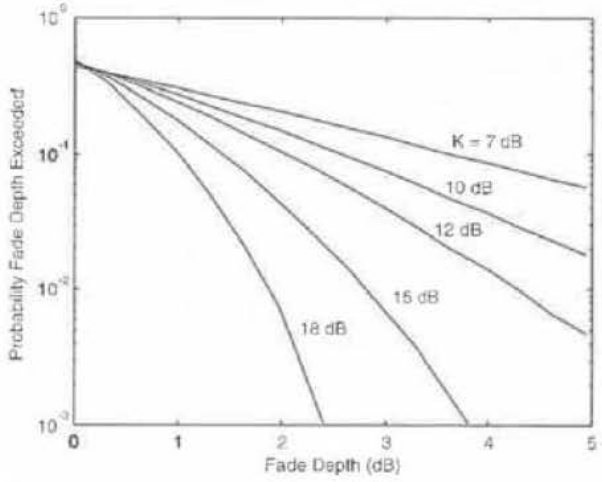


FIGURE 74.3 Cumulative fade distribution for the Rician channel as a function of K factor.

and the corresponding time correlation of the diffuse component is given by

$$R(\tau) = J_0(2\pi f_d \tau) \tag{74.7}$$

where $J_0(\cdot)$ is the zeroth-order Bessel function.

For the land–mobile channel, some effort has been devoted to providing an appropriate model combining the effects of shadowing and multipath fading, which is appropriate for satellite applications. An example of these is the Lutz model [Lutz et al., 1991], which consists of two states, a good state and a bad state. In the good state, the channel is Rician. In the bad state, the line-of-sight signal is essentially blocked due to significant shadowing. The parameters of the model are the distribution of the dwell times in each state, the K factor, and the distribution of the shadowing, which are selected to model the environment of interest.

In aeronautical–mobile, the fading spectrum is generally modeled as Gaussian. However, measurements in this case indicate that the fading bandwidth is not strongly dependent on the aircraft velocity. The fading tends to be slow and is due to reflections from a limited portion of the Earth's surface and from the aircraft's body. At L-band frequencies (1.6 GHz) the fading bandwidths for aeronautical–mobile channels range from 20 to 100 Hz with 20 Hz being more prevalent [Neul et al., 1987].

The physics involved in determining an appropriate model for the maritime channel is quite similar to that for the aeronautical case [Moreland, 1987]. Consequently, the fading spectrum is generally modeled as Gaussian. However, the fading bandwidth is typically less than 1 Hz [Schweikert and Hagenauer, 1983].

In general, multipath in satellite–mobile channels is not time dispersive. Measurements indicate that for practical bandwidths (<2 MHz), this is true even for wideband channels [Jahn et al., 1995]. A possible exception to this is the high-altitude aeronautical channel where one may occasionally receive distant reflections from the sea surface or ice. For this case, the strength of the reflected path is strong enough to be problematic only at elevation angles below 20°.

74.3 Satellite Orbits

Initial mobile satellite systems used geostationary orbits and were generally aimed at mobile terminals with significant antenna gain (>8 dB). These include a number of regional mobile satellite

Samsung Exhibit 1022, Page 136 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

85.2 Cellular Systems and Power Control

In cellular communication systems, the service area is divided into cells, each covered by a single base station. If, in the forward link (base station to mobile), all users served by all base stations share the same frequency, each communication between a base station and a particular user would also reach all other users in the form of cochannel interference. The greater the distance between the mobile and the interfering transmitter, however, the weaker the interference becomes due to the propagation loss. To ensure a good quality of service throughout the cell, the received signal in the fringe area of the cell must be strong. Once the signal has crossed the boundary of a cell, however, it becomes interference and is required to be as weak as possible. Since this is difficult, the channel frequency is usually not reused in adjacent cells in most of the cellular systems. If the frequency is reused, the cochannel interference damages the signal reception in the adjacent cell, and the quality of service severely degrades unless other measures are taken to mitigate the interference. Therefore, a typical reuse pattern reuses the frequency in every seventh cell (frequency reuse factor = 1/7). The only exception is for CDMA-based systems where the users are separated by codes, and the allocated frequency may be shared by all users in all cells.

Even if the frequency is reused in every seventh cell, there is still some cochannel interference arriving at the receiver. It is, therefore, very important to maintain a minimal transmitted level at the base station to keep the cochannel interference low, frequency reuse factor high, and therefore the capacity of the system and quality of service high.

The same principle applies in the reverse link (mobile to base station); the power control maintains the minimum necessary transmitted power for reliable communication. Several additional benefits can be gained from this strategy. The lower transmitted power conserves the battery energy allowing the mobile terminal (the portable) to be lighter and stay on the air longer. Furthermore, recent concerns about health hazards caused by the portable's electromagnetic emissions are also alleviated.

In the reverse link, the power control also serves to alleviate the near–far effect. If all mobiles transmitted at the same power level, the signal from a near mobile would be received as the strongest. The difference between the received signal strength from the nearest and the farthest mobile can be in the range of 100 dB, which would cause saturation of the weaker signals' receivers or an excessive amount of adjacent channel interference. To avoid this, the transmitted power at the mobile must be adjusted inversely proportional to the effective distance from the base station. The term effective distance is used since a closely located user in a propagation shadow or in a deep fade may have a weaker signal than a more distant user having excellent propagation conditions.

In a DS/CDMA system, power control is a vital necessity for system operation. The capacity of a DS/CDMA cellular system is interference limited since the channels are neither separated in frequency nor separated in time, and the cochannel interference is inherently strong. A single user exceeding the limit on transmitted power could inhibit the communication of all other users.

The power control systems have to compensate not only for signal strength variations due to the varying distance between base station and mobile but must also attempt to compensate for signal strength fluctuations typical of a wireless channel. These fluctuations are due to the changing propagation environment between the base station and the user as the user moves across the cell or as some elements in the cell move. There are two main groups of channel fluctuations: slow (i.e., **shadowing**) and fast **fading**.

As the user moves away from the base station, the received signal becomes weaker because of the growing propagation attenuation with the distance. As the mobile moves in uneven terrain it often travels into a propagation shadow behind a building or a hill or other obstacle much larger than the wavelength of the frequency of the wireless channel. This phenomenon is called shadowing. Shadowing in a land-mobile channel is usually described as a stochastic process having log-normal distributed amplitude. For other types of channels other distributions are used, e.g., Nakagami.

Samsung Exhibit 1022, Page 137 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Electromagnetic waves transmitted from the transmitter may follow multiple paths on the way from the transmitter to the receiver. The different paths have different delays and interfere at the antenna of the receiver. If two paths have the same propagation attenuation and their delay differs in an odd number of half-wavelengths (half-periods), the two waves may cancel each other at the antenna completely. If the delay is an even multiple of the half-wavelengths (half-periods), the two waves may constructively add, resulting in a signal of double amplitude. In all other cases (nonequal gains, delays not a multiple of half-wavelength), the resultant signal at the antenna of the receiver is between the two mentioned limiting cases. This fluctuation of the channel gain is called fading. Since the scattering and reflecting surfaces in the service area are randomly distributed (houses, trees, furniture, walls, etc.), the amplitude of the resulting signal is also a random variable. The amplitude of fading is usually described by a Rayleigh, Rice, or Nakagami distributed random variable.

Since the mobile terminal may move at the velocity of a moving car or even of a fast train, the rate of channel fluctuations may be quite high and the power control has to react very quickly in order to compensate for it. The rate of fading is usually expressed in terms of Doppler frequency.

The performance of the **reverse link** of DS/CDMA systems is most affected by the near–far effect and, therefore, very sophisticated power control systems in the reverse link that attempt to alleviate the effects of channel fluctuations must be used. Together with other techniques, such as micro- and macrodiversity, interleaving, and coding, the DS/CDMA cellular system is able to cope with the wireless channel extremely well.

The effective use of the **power control** in DS/CDMA cellular system enables the frequency to be reused in every cell, which in turn enables features such as the soft hand-off and base station diversity. All together, these help enhance the capacity of the system.

In the **forward link** of a DS/CDMA system, power control may also be used. It may vary the transmitted power to the mobile, but the dynamic range is smaller due to the shared spectrum and, thus, shared interference.

We can distinguish between two kinds of power control, the open-loop power control and the closed-loop power control. The open-loop power control estimates the channel and adjusts the transmitted power accordingly but does not attempt to obtain feedback information on its effectiveness. Obviously, the open-loop power control is not very accurate, but since it does not have to wait for the feedback information it may be relatively fast. This can be advantageous, in the case of a sudden channel fluctuation, such as a mobile driving from behind a big building. This fast action is required, for instance, in the reverse link of a DS/CDMA system where the sudden increase of received strength at the base station may suppress all other signals.

The principle operation of open-loop power control is shown in Fig. 85.1. The open-loop power control must base its action on the estimation of the channel state. In the reverse link it estimates the channel by measuring the received power level of the pilot from the base station in the forward link and sets the transmitted power level inversely proportional to it. Ideally, this ensures that the average power level received from the mobile at the base station remains constant irrespective of the channel variations. This approach, however, assumes that the forward and the reverse link signal strengths are closely correlated. Although forward and reverse link may not share the same frequency and, therefore, the fading is significantly different, the long-term channel fluctuations due to shadowing and propagation loss are basically the same.

The closed-loop power control system [Fig. 85.2(a)] may base its decision on an actual communication link performance metric, e.g., received signal power level, received signal-to-noise ratio, received bit-error rate, or received frame-error rate. In the case of the reverse link power control, this metric may be forwarded to the mobile as a base for an autonomous power control decision, or the metric may be evaluated at the base station and only a power control adjustment command is transmitted to the mobile. If the reverse link power control decision is made at the base station, it

Samsung Exhibit 1022, Page 138 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

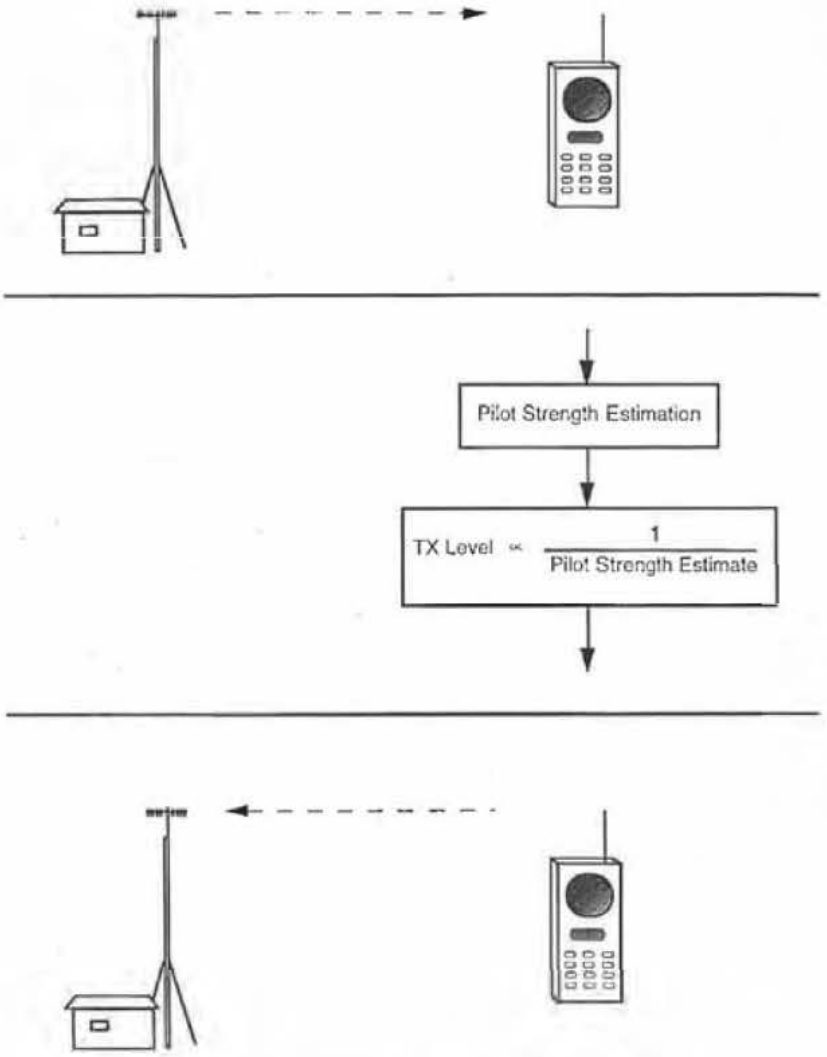


FIGURE 85.1 Reverse link open-loop power control.

may be based on the additional knowledge of the particular mobile's performance and/or a group of mobiles' performance (such as mobiles in a sector, cell, or even in a cluster of cells). If the power control decision for a particular mobile is made at the base station or at the switching office for all mobiles and is based on the knowledge of all other mobile's performance, it is called a centralized power control system. A centralized power control system may be more accurate than a distributed power control system, but it is much more complex in design, more costly, and technologically challenging.

In principle, the same categorization may be used for the power control in the forward link [Fig. 85.2(b)] except that in the forward link pilots from the mobiles are usually unavailable and only closed-loop power control is applied.

In the ideal case, power control compensates for the propagation loss, shadowing, and fast fading. There are many effects, however, that prevent the power control from becoming ideal. Fast fading rate, finite delays of the power control system, nonideal channel estimation, error in the power control command transmission, limited dynamic range, etc., all contribute to degrading the performance of the power control system. It is very important to examine the performance of power

Samsung Exhibit 1022, Page 139 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

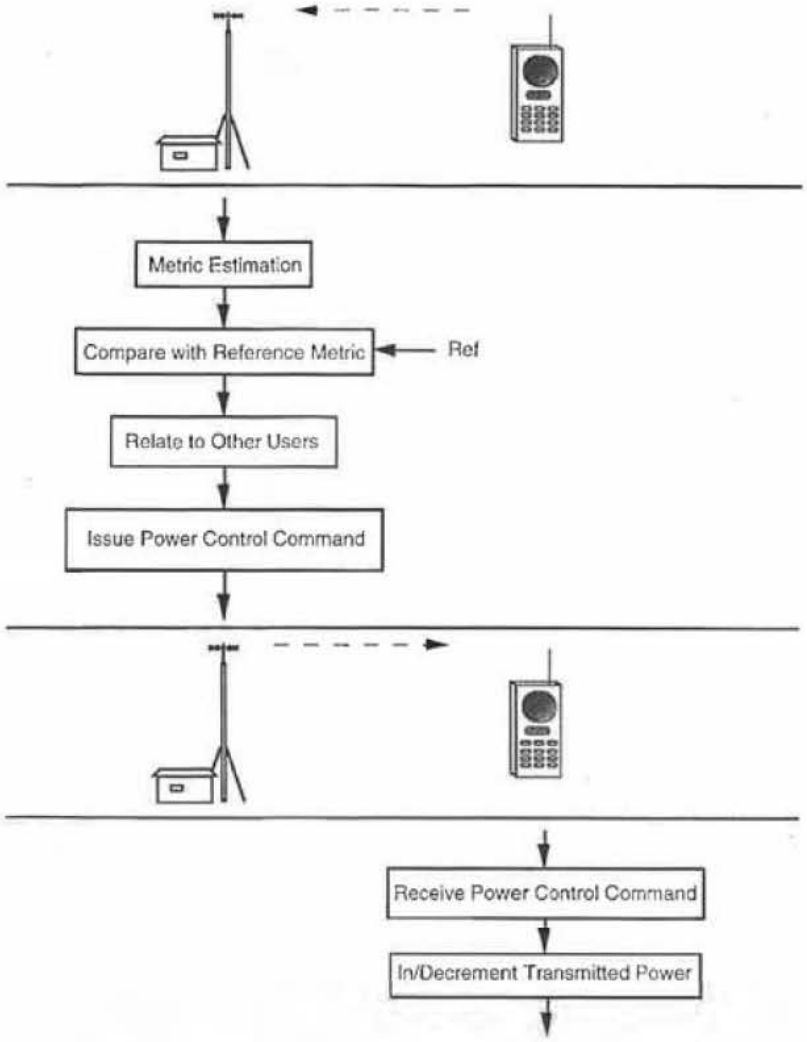


FIGURE 85.2(a) Reverse link closed-loop power control.

control under nonideal conditions since the research done has shown that the power control system is quite sensitive to some of these conditions [Viterbi and Zehavi, 1993]. Kudoh, 1993, simulated a nonideal closed-loop power control system. Errors in the system were represented by a log-normal distributed control error with standard deviation σ_E in decibels. Some results on capacity reduction are presented in Table 85.1.

The effects of Doppler and delay and feedback errors in power control loop on power control have also been studied. [Pichna et al., 1993].

TABLE 85.1 Capacity Reduction Versus Power Control Error

	$\sigma_E = 0.5$ dB, %	$\sigma_E = 1$ dB, %	$\sigma_E = 2$ dB, %	$\sigma_E = 3$ dB, %
Forward link	10	29	64	83
Reverse link	10	31	61	81

Source: Kudoh, E. 1993. On the capacity of DS/CDMA cellular mobile radios under imperfect transmitter power control. *IEICE Trans. Commun.* E76-B(April):886–893.

Samsung Exhibit 1022, Page 140 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

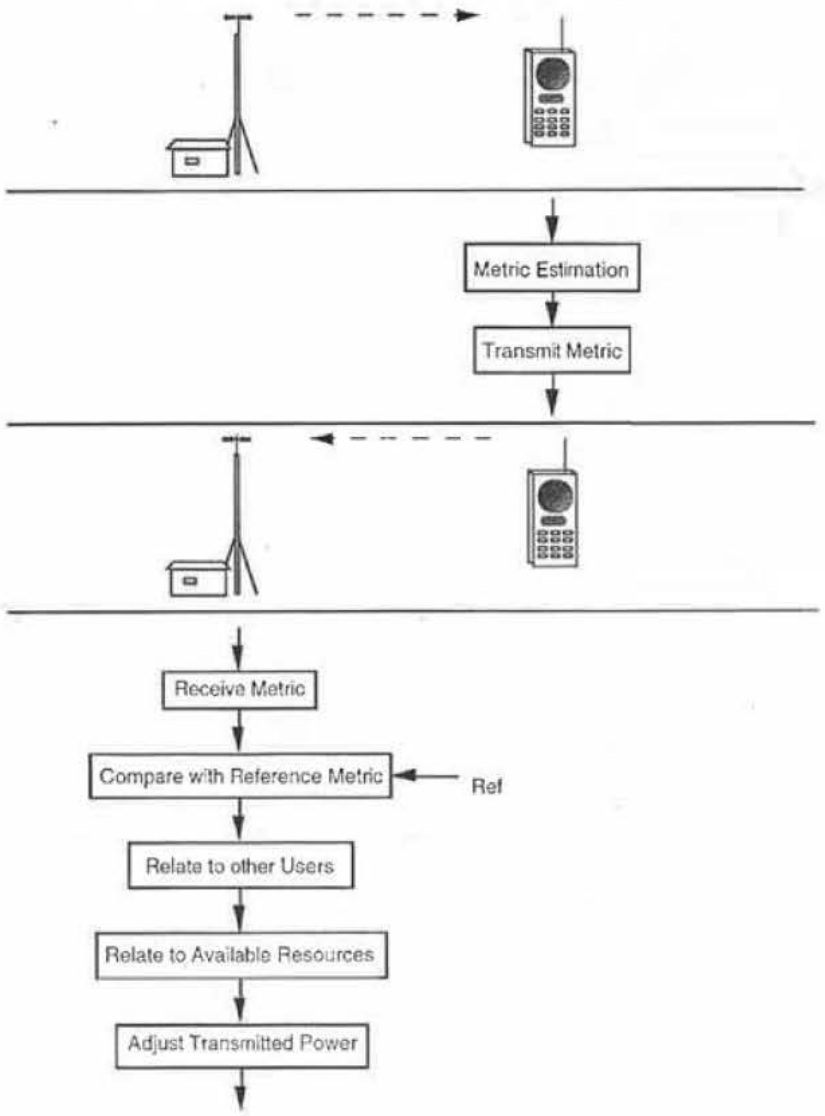


FIGURE 85.2(b) Forward link closed-loop power control.

85.3 Power Control Examples

In the following section, several applications of power control of analog and digital cellular systems are presented.

In the analog networks we may see power control implemented in both the reverse link and forward link [Lee, 1989]. Power control in the reverse link 1) reduces the chance of receiver saturation by a closely located mobile, 2) reduces the co-channel interference and thus increases the frequency-reuse factor and capacity, and 3) reduces the average transmitted power at the mobile thus conserving battery energy at the mobile.

The power control in the forward link 1) reduces co-channel interference and thus increases the frequency reuse factor and capacity and 2) reduces adjacent-channel interference and improves the quality of service.

One example of a power control system shown by Lee, 1993, was of an air-to-ground communication system. The relevant airspace is divided into six zones based on the aircraft altitude. The transmitted power at the aircraft is then varied in six steps based on the zone in which the aircraft is

Samsung Exhibit 1022, Page 141 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Similarly, $P1$ is the energy contribution of the code vector selected from the first codebook as a fraction of the total excitation energy of the subframe. The transformation reduces the dynamic range of the parameters to be encoded. An 8-b vector quantizer efficiently encodes the appropriate (GS, $P0$, $P1$) vectors by selecting the vector which minimizes the weighted error. The received and decoded values β, γ_1, and γ_2 are computed from the received (GS, $P0$, $P1$) vector and applied to reconstitute the decoded signal.

88.13 Channel Coding and Interleaving

The goals of channel coding are to reduce the impairments in the reconstructed speech due to transmission errors. The 159 b characterizing each 20-ms block of speech are divided into two classes, 77 in class 1 and 82 in class 2. Class 1 includes the bits in which errors result in a more significant impairment, whereas the speech quality is considered less sensitive to the class- 2 bits. Class 1 generally includes the gain, pitch lag, and more significant reflection coefficient bits. In addition, a 7-b cyclic redundancy check is applied to the 12 most perceptually significant bits of class 1 to indicate whether the error correction was successful. Failure of the CRC check at the receiver suggests that the received information is so erroneous that it would be better to discard it than use it. The error correction coding is illustrated in Fig. 88.5.

The error correction technique used is rate 1/2 convolutional coding with a constraint length of 5 [Lin and Costello, 1983]. A tail of 5 b is appended to the 84 b to be convolutionally encoded to result in a 178-b output. Inclusion of the tail bits ensures independent decoding of successive time slots and no propagation of errors between slots.

Interleaving the bits to be transmitted over two time slots is introduced to diminish the effects of short deep fades and to improve the error-correction capabilities of the channel coding technique. Two speech frames, the previous and the present, are interleaved so that the bits from each speech block span two transmission time slots separated by 20 ms. The interleaving attempts to separate

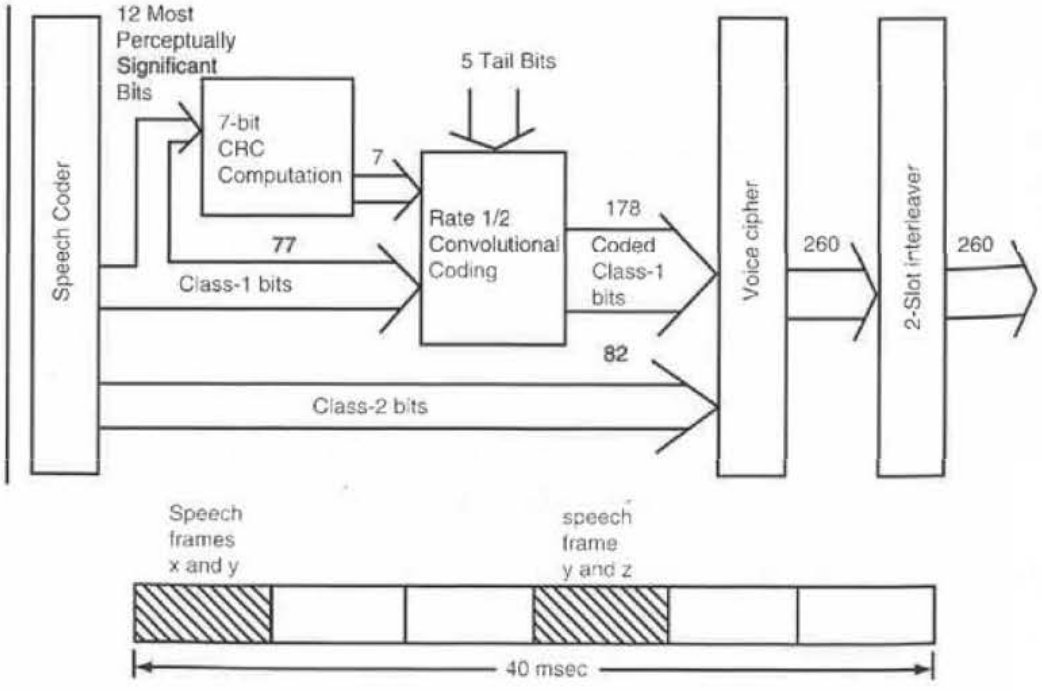


FIGURE 88.5 Error correction insertion for speech coder. Source TIA, 1992. Cellular Systems Dual-Mode Mobile Station–Base Station Compatibility Standards. TIA/EIA IS-54. With permission.

Samsung Exhibit 1022, Page 142 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

the convolutionally coded class-1 bits from one frame as much as possible in time by inserting noncoded class-2 bits between them.

88.14 Bad Frame Masking

A CRC failure indicates that the received data is unusable, either due to transmission errors resulting from a fade, or from pre-emption of the time slot by a control message (fast associated control channel, FACCH). To mask the effects that may result from leaving a gap in the speech signal, a masking operation based on the temporal redundancy between adjacent speech blocks has been proposed. Such masking can at best bridge over short gaps but cannot recover loss of signal of longer duration. The bad frame masking operation may follow a finite state machine where each state indicates an operation appropriate to the elapsed duration of the fade to which it corresponds. The masking operation consists of copying the previous LPC information and attenuating the gain of the signal. State 6 corresponds to error sequences exceeding 100 ms, for which the output signal is muted. The result of such a masking operation is generation of an extrapolation in the gap to the previously received signal, significantly reducing the perceptual effects of short fades. No additional delay is introduced in the reconstructed signal. At the same time, the receiver will report a high frequency of bad frames leading the system to explore handoff possibilities immediately. A quick successful handoff will result in rapid signal recovery.

88.15 Conclusions

The IS-54 digital cellular standard specifies modulation and speech coding techniques for mobile cellular systems that allow the interoperation of terminals built by a variety of manufacturers and systems operated across the country by a number of different service providers. It permits speech communication with good quality in a transmission environment characterized by frequent multipath fading and significant intercell interference. Generally, the quality of the IS-54 decoded speech is better at the edges of a cell than the corresponding AMPS transmission due to the error mitigation resulting from channel coding. Near a base station or in the absence of significant fading and interference, the IS-54 speech quality is reported to be somewhat worse than AMPS due to the inherent limitations of the analysis–synthesis model in reconstructing arbitrary speech signals with limited bits. At this time the TIA is considering introducing an alternative higher quality 8-kb/s speech coding technique for applications where higher speech quality is desirable. The International Telecommunications Union (ITU) is in the process of selecting such a coding algorithm, which promises speech quality equivalent to wireline telephony links in the absence of transmission errors [Salami et al., 1994]. The selection of the standard placed significant emphasis on the ability to implement the transceiver and encoder/decoder in a variety of mobile and portable terminals at low cost and with low-power dissipation. The standard selected reflects a practically reasonable compromise between high performance and low complexity. The low-weight terminals permitting hours of talk time in mobile environments attest to the benefits of the compromises arrived at in the definition of the standard.

Defining Terms

Codebook: A set of signal vectors available to both the encoder and decoder.

Covariance lattice algorithm: An algorithm for reduction of the covariance matrix of the signal consisting of several lattice stages, each stage implementing an optimal first-order filter with a single coefficient.

Reflection coefficient: A parameter of each stage of the lattice linear prediction filter that determines 1) a forward residual signal at the output of the filter-stage by subtracting from the forward residual at the input a linear function of the backward residual, also 2) a backward

Samsung Exhibit 1022, Page 143 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

LLME

All of the network layer entities interface to the LLME which provides coordination of the operation between different network layer entities and also between the network and lower layers. The LLME interfaces with the physical, MAC, DLC, and NWK layers to provide procedures that concern more than one layer. Most of these procedures have only local significance.

LLME Physical Layer: PM-SAP

The primitives passed through PM-SAP are mainly used to invoke and control physical layer processes. The following primitives are exchanged through PM-SAP: PL-ME-SYNC, PL-ME-SIG_STR, and PL-ME-TIME_ADJ. Using these primitives the LLME implements procedures to produce: the list of quietest channels, the list of physical channels with greatest field strength (PP only), and timing information.

Prior to the use of a DECT physical channel, the receiver measures the strength of signals on that physical channel using the PL-ME_STR primitive. Using these signal measurements, the LLME produces two ordered lists: a list of least interfered channels and a list channels with greatest field strength (PP only). Using the PL-ME-SYNC primitives and interworking with higher layer detection of slot numbers, the LLME extracts timing information to establish the slot and frame timing.

MAC Layer

- Creation, maintenance and release of bearers, by activating and deactivating pairs of physical channels.
- Physical channels management, including the choice of free physical channels and the assessment of quality of received signals.

Connection and control.

MAC-ME-CON: Connection setup
MAC-ME-CON_ALL: Connection setup allowed
MAC-ME-REL: Bearer release
MAC-ME-REL_REP: MBC release support

System information and identities.

MAC-ME-RFP_PRELOAD: FP information preloading
MAC-ME-PT_PRELOAD: PT information preloading
MAC-ME-INFO: Systems information output
MAC-ME-EXT: Extended system information
MAC-ME-CHANMAP: Channel map
MAC-ME-STATUS: Status report
MAC-ME-ERROR: Error reports

Procedures.

- Broadcasting
- Extended system information
- PP states and transitions
- Physical channel selection
- In-connection quality control
- Maximum allowed system load at RFPs

Samsung Exhibit 1022, Page 144 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

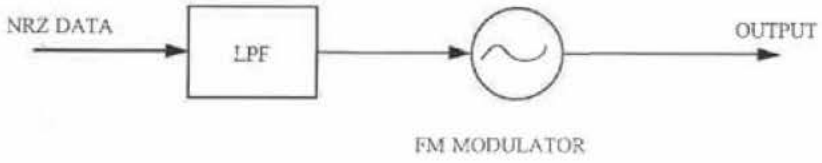


FIGURE 92.10 Premodulation baseband-filtered MSK.

- PMID and FMID definitions
- RFP idle receiver scan sequence
- PT fast setup receiver scan sequence

Data Link Layer (DLC)

DLC covers routing of C-plane and U-plane data to suitable connections and connection management, which includes the establishment and release of connections in response to network layer demands. It also provides coordination and control for the C-plane and U-plane processes. The LLME controls the routing of the C-plane and U-plane frames from the available MAC connections and controls the opening and closing (handover) of the MAC connections in response to service demands.

Network Layer

The network layer provides service negotiation and mapping. All of the network layer entities interface to the LLME, which provides coordination of the operation between different network layer entities and also between the network and lower layers.

Modulation Method

The modulation method for DECT is Gaussian filtered frequency shift keying (GFSK) with a nominal deviation of 288 kHz [Madsen and Fague, 1993]. The BT, i.e., Gaussian filter bandwidth to bit ratio, is 0.5 and the bit rate is 1.152 Mb/s. Specification details can be obtained from the relevant ETSI documents listed in the reference section.

Digital transmission channels in the radio frequency bands, including the DECT systems, present serious problems of spectral congestion and introduce severe adjacent/cochannel interference problems. There were several schemes employed to alleviate these problems: new allocations at high frequencies, use of frequency-reuse techniques, efficient source encoding, and spectrally efficient modulation techniques.

Any communication system is governed by mainly two criteria, transmitted power and channel bandwidth. These two variables have to be exploited in an optimum manner in order to achieve maximum bandwidth efficiency, defined as the ratio of data rate to channel bandwidth (units of bits/Hz/s) [Pasupathy, 1979]. GMSK/GFSK has the properties of constant envelope, relatively narrow bandwidth, and coherent detection capability. Minimum shift keying (MSK) can be generated directly from FM, i.e., the output power spectrum of MSK can be created by using a premodulation low-pass filter. To ensure that the output power spectrum is constant, the low-pass filter should have a narrow bandwidth and sharp cutoff, low overshoot, and the filter output should have a phase shift $\pi/2$, which is useful for coherent detection of MSK; see Fig. 92.10.

Properties of GMSK satisfy all of these characteristics. We replace the low-pass filter with a premodulation Gaussian low-pass filter [Murota and Hirade, 1981]. As shown in Fig. 92.11, it is relatively simple to modulate the frequency of the VCO directly by the baseband Gaussian pulse stream, however, the difficulty lies in keeping the center frequency within the allowable value. This becomes more apparent when analog techniques are employed for generating such signals. A possible solution to this problem in the analog domain would be to use a phase-lock loop (PLL)

Samsung Exhibit 1022, Page 145 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

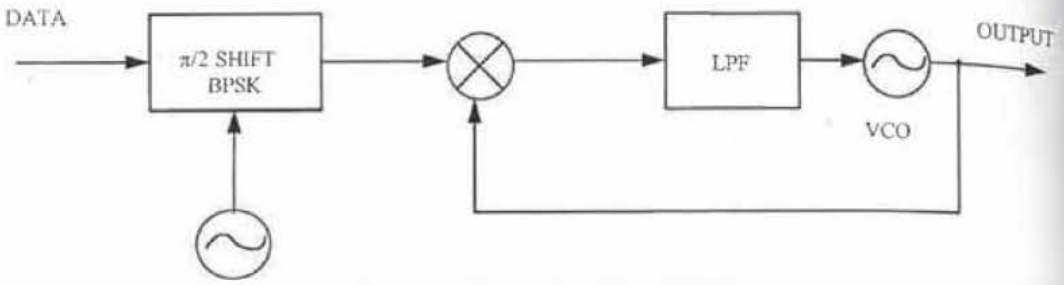


FIGURE 92.11 PLL type GMSK modulator.

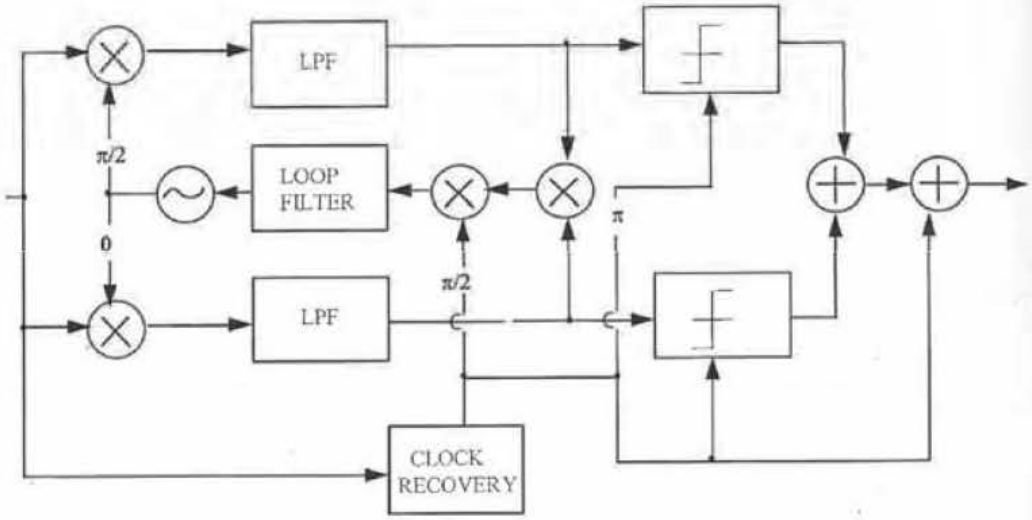


FIGURE 92.12 Costas Loop.

modulator with a precise transfer function. It is desirable these days to employ digital techniques, which are far more robust in meeting the requirements talked about earlier. This would suggest an orthogonal modulator with digital waveform generators [de Jager and Dekker, 1978].

The demodulator structure in a GMSK/GFSK system is centered around orthogonal coherent detection, the main issue being recovery of the reference carrier and timing. A typical method, is described in de Buda, 1972, where the reference carrier is recovered by dividing by four the sum of the two discrete frequencies contained in the frequency doubler output, and the timing is recovered directly from their difference. This method can also be considered to be equivalent to the Costas loop structure as shown in Fig. 92.12.

In the following are some theoretical and experimental representations of the modulation technique just described. Considerable literature is available on the subject of data and modulation schemes and the reader is advised to refer to Pasupathy, 1979, and Murota and Hirade, 1981, for further access to relevant study material.

Radio Frequency Architecture

We have discussed the need for low power consumption and low cost in designing cordless telephones. These days digital transmitter/single conversion receiver techniques are employed to provide highly accurate quadrature modulation formats and quadrature down conversion schemes that allow a great deal of flexibility to the baseband section. Generally, one would have used digital

Samsung Exhibit 1022, Page 146 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

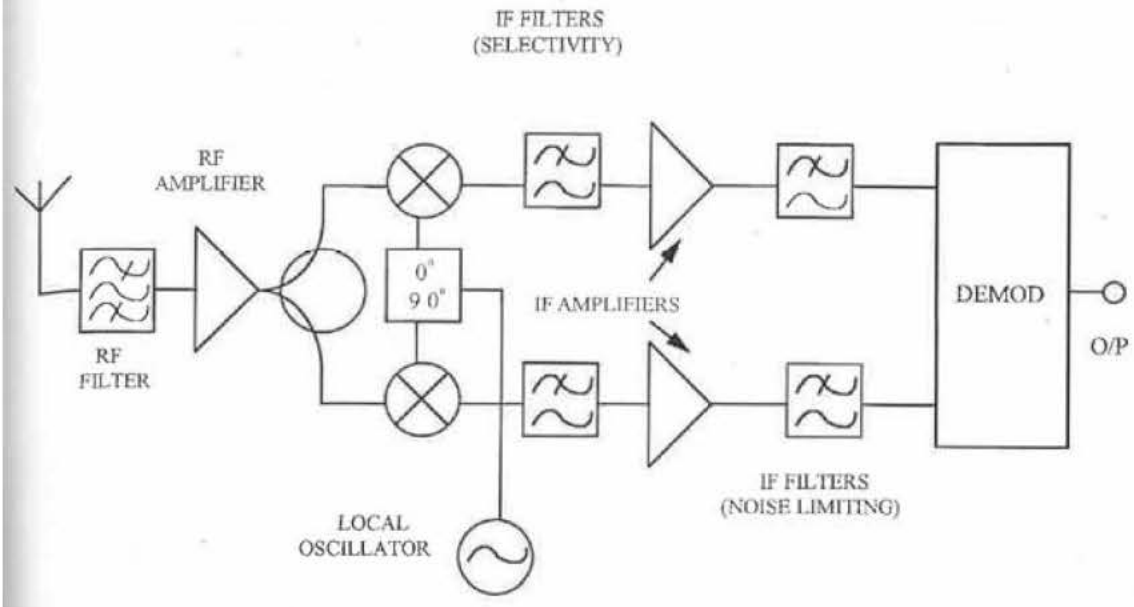


FIGURE 92.13 Direct conversion receiver architecture.

signal processors to perform most of the demodulation functions at the cost of high current consumption. With the advent of application specific signal processing, solutions with these techniques have become more attractive.

From a system perspective, range, multipath, and voice quality influence the design of a DECT phone. A high bit rate coupled with multipath reflections in an indoor environment makes DECT design a challenging task. The delay spread (multipath) can be anywhere in the 100–200 ns range, and a DECT bit time is 880 ns. Therefore, a potential delay spread due to multipath reflections is 1–20% of a bit time. Typically, antenna diversity is used to overcome such effects.

DECT employs a TDMA/TDD method for transmission, which simplifies the complexity of the radio frequency end. The transmitter is on for 380 ms or so. The receiver is also only on for a similar length of time.

A single conversion radio architecture requires fast synthesizer switching speed in order to transmit and receive on as many as 24 timeslots per frame. In this single conversion transmitter structure, the synthesizer has to make a large jump in frequency between transmitting and receiving, typically in the order of 110 MHz. For a DECT transceiver, the PLL synthesizer must have a wide tuning bandwidth at a high-frequency reference in addition to good noise performance and fast switching speed. The prescaler and PLL must consume as low a current as possible to preserve battery life.

In the receive mode the RF signal at the antenna is filtered with a low-loss antenna filter to reduce out-of-band interfering signals. This filter is also used on the transmit side to attenuate harmonics and reduce wideband noise. The signal is further filtered, shaped, and down converted as shown in the Fig. 92.13. The signal path really is no different from most receiver structures. The challenges lie in the implementation, and this area has become quite a competitive segment, especially in the semiconductor world.

The direct conversion receiver usually has an intermediate frequency nominally at zero frequency, hence the term zero IF. The effect of this is to fold the spectrum about zero frequency, which results in the signal occupying only one-half the bandwidth. The zero IF architecture possesses several advantages over the normal superheterodyne approach. First, selectivity requirements for the RF filter are greatly reduced due to the fact that the IF is at zero frequency and the image response is coincident with the wanted signal frequency. Second, the choice of zero frequency means that the

Samsung Exhibit 1022, Page 147 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

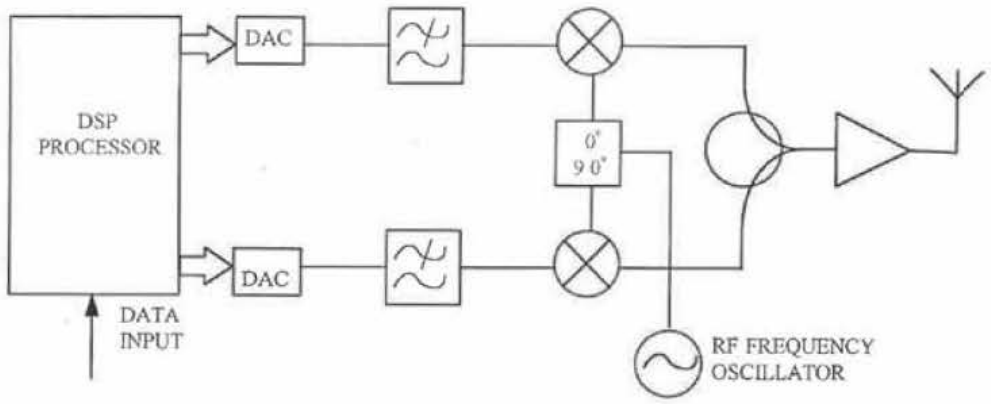


FIGURE 92.14 Transmit section.

bandwidth for the IF paths is only half the wanted signal bandwidth. Third, channel selectivity can be performed simply by a pair of low-bandwidth low-pass filters.

For the twin IF chains of a direct conversion receiver, automatic gain control (AGC) is always required due the fact that each IF channel can vary between zero and the envelope peak at much lower rates than the highest signal bandwidth frequency. An additional requirement in newer systems is received signal strength indication (RSSI) to measure the signal or interference level on any given channel.

Moving on to the transmitter architecture (shown in Fig. 92.14 is a typical I-Q system), it is safe to say that the task of generating an RF signal is much simpler than receiving it. A transmitter consists of three main components: a final frequency generator, a modulator, and the power amplifier. These components can all be combined in common circuits, i.e., frequency synthesizer with inbuilt modulator. The problem of generating a carrier at a high frequency is largely one of frequency control. The main approach for accurately generating an output frequency from a crystal reference today is the PLL, and there is considerable literature available on the subject [Gardner, 1979]. In the modulation stage, depending upon the tightness of the phase accuracy specification of a cordless system, it may be necessary to apply tight control on the modulation index to ensure that the phase path of the signal jumps exactly in 90° increments. Description of an I-Q modulator is also given in a previous chapter.

Defining Terms

AGC: Automatic gain control
ARQ: Automatic repeat request
AWGN: Additive white Gaussian noise
BABT: British approvals board for telecommunications
Base Station: The fixed radio component of a cordless link. This may be single-channel (for domestic) or multi-channel (for Telepoint and business)
BER: Bit error rate (or ratio)
CCITT: Comitè Consultatif International des Tèlègraphes et Tèlèphones, part of the ITU
CEPT: Conference of European Posts and Telecommunications Administrations
CPFSK: Continuous phase frequency shift keying
CPP: Cordless portable part, the cordless telephone handset carried by the user
CRC: Cyclic redundancy check
CT2: Second generation cordless telephone-digital
D Channel: Control and information data channel (16 kb/s in ISDN)

Samsung Exhibit 1022, Page 148 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

DCT: Digital cordless telephone
DECT: Digital European cordless telecommunications
DLC: Data link control layer, protocol layer in DECT
DSP: Digital signal processing
DTMF: Dual tone multiple frequency (audio tone signalling system)
ETSI: European Telecommunications Standards Institute
FDMA: Frequency division multiple access
FSK: Frequency shift keying
GMSK: Gaussian filtered minimum shift keying
ISDN: Integrated services digital network
ITU: International Telecommunications Union
MPT 1375: U.K. standard for common air interface (CAI) digital cordless telephones
MSK: Minimum shift keying
PSK: Phase shift keying
RES 3: Technical subcommittee, radio equipment and systems 3 of ETSI, responsible for the specification of DECT
RSSI: Received signal strength indication
SAW: Surface acoustic wave
TDD: Time division duplex
TDMA: Time division multiple access

References

Cheer, A.P. 1985. Architectures for digitally implemented radios, IEE Colloquium on Digitally Implemented Radios, London.

Comité Consultatif International des Télégraphes et Téléphones. 1984, "32 kbits/sec Adaptive Differential Pulse Code Modulation (ADPCM)," CCITT Red Book, Fascicle III.3, Rec G721.

Comité Consultatif International des Télégraphes et Téléphones. 1984–1985. *General Characteristics of International Telephone Connections and Circuits*, CCITT Red Book, Vol. 3, Facicle III.1, Rec. G101–G181.

de Buda, R. 1972. Coherent demodulation of frequency shifting with low deviation ratio, *IEEE Trans.* COM-20 (June):466–470.

de Jager, F. and Dekker, C.B. 1978. Tamed frequency modulation. A novel method to achieve spectrum ecomony in digital transmission, *IEEE Trans. in Comm.* COM-20 (May):534–542.

Dijkstra, S. and Owen, F. 1994. The case for DECT, *Mobile Comms Int.* :60–65.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-1 (Overview). Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-2 (Physical Layer) Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-3 (MAC Layer) Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-4 (Data Link Control Layer) Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-5 (Network Layer) Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-6 (Identities and Addressing) Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-7 (Security Features) Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

European Telecommunications Standards Inst. 1992. RES-3 DECT Ref. Doc. ETS 300 175-8 (Speech Coding & Transmission) Oct. ETSI Secretariat, Sophia Antipolis Ceder, France.

Samsung Exhibit 1022, Page 149 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

The CELP framework has evolved to form the basis of a great variety of speech coding algorithms, including all existing full- and half-rate standard algorithms for digital cellular systems. We outline next the basic CELP encoder-processing steps, in a form suited to our subsequent detailed descriptions of the PDC and GSM half-rate coders. These steps have accounted for various computational efficiency considerations and may, therefore, deviate from a conceptual functional description of the encoder constituents.

1. LP analysis on the current frame of input speech to determine the coefficients of the all-pole synthesis filter;
2. quantization of the LP filter parameters;
3. determination of the open-loop **pitch period** or lag;
4. adapting the perceptual weighting filter to the current LP information (and also pitch information when appropriate) and applying the adapted filter to the input speech signal;
5. formation of a filter cascade (which we shall refer to as *perceptually weighted synthesis filter*) consisting of the LP synthesis filter, as specified by the quantized parameters in step 2, followed by the perceptual weighting filter;
6. subtraction of the *zero-input response* of the perceptually weighted synthesis filter (the filter's decaying response due to past input) from the perceptually weighted input speech signal obtained in step 4;
7. an *adaptive codebook* is searched to find the most suitable periodic excitation, i.e., when the perceptually weighted synthesis filter is driven by the best code vector from the adaptive codebook, the output of the filter cascade should best match the difference signal obtained in step 6;
8. one or more nonadaptive excitation codebooks are searched to find the most suitable random excitation vectors that, when added to the best periodic excitation as determined in step 7 and with the resultant sum signal driving the filter cascade, would result in an output signal best matching the difference signal obtained in step 6.

Steps 1–6 are executed once per frame. Steps 7 and 8 are executed once for each of the *subframes* that together constitute a frame. Step 7 may be skipped depending on the pitch information from step 3, or if step 7 were always executed, a *nonperiodic excitation* decision would be one of the possible outcomes of the search process in step 7. Integral to steps 7 and 8 is the determination of gain (scaling) parameters for the excitation vectors. For each frame of input speech, the filter and excitation and gain parameters determined as outlined are conveyed as encoded bits to the speech decoder.

In a properly designed system, the data conveyed by the channel decoder to the speech decoder should be free of errors most of the time, and the speech signal synthesized by the speech decoder would be identical to that as determined in the speech encoder's ABS operation. It is common to enhance the quality of the synthesized speech by using an adaptive *postfilter* to attenuate coder noise in the perceptually sensitive regions of the spectrum. The postfilter of the decoder and the perceptual weighting filter of the encoder may seem to be functionally identical. The weighting filter, however, influences the selection of the best excitation among available choices, whereas the postfilter actually shapes the spectrum of the synthesized signal. Since postfiltering introduces its own distortion, its advantage may be diminished if tandem coding occurs along the end-to-end communication path. Nevertheless, proper design can ensure that the net effect of postfiltering is a reduction in the amount of audible codec noise [Chen and Gersho, 1995]. Excepting postfiltering, all other speech synthesis operations of an LPAS decoder are (effectively) duplicated in the encoder (though the converse is not true). Using this fact, we shall illustrate each coder in the sequel by exhibiting only a block diagram of its encoder or decoder but not both.

94.5 Channel Coding Techniques in the Half-Rate Standards

Crucial to the maintainence of quality speech communication is the ability to transport coded speech data across the radio channel with minimal errors. Low-bit-rate LPAS coders are particularly

Samsung Exhibit 1022, Page 150 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

sensitive to channel errors; errors in the bits representing the LP parameters in one frame, for instance, could result in the synthesis of nonsensical sounds for longer than a frame duration. The error rate of a digital cellular radio channel with no channel coding can be catastropically high for LPAS coders. The amount of tolerable transmission delay is limited by the requirement of interactive communication and, consequently, *forward error control* is used to remedy transmission errors. "Forward" means that channel errors are remedied in the receiver, with no additional information from the transmitter and, hence, no additional transmission delay. To enable the channel decoder to correct channel errors, the channel encoder conveys more bits than the amount generated by the speech encoder. The additional bits are for error *protection*, as errors may or may not occur in any particular transmission epoch. The ratio of the number of encoder input (information) bits to the number of encoder output (code) bits is called the (channel) *coding rate*. This is a number no more than one and generally decreases as the error protection power increases. Though a lower channel coding rate gives more error protection, fewer bits will be available for speech coding. When the channel is in good condition and, hence, less error protection is needed, the received speech quality could be better if bits devoted to channel coding were used for speech coding. On the other hand, if a high channel coding rate were used, there would be uncorrected errors under poor channel conditions and speech quality would suffer. Thus, when nonadaptive forward error protection is used over channels with nonstationary statistics, there is an inevitable tradeoff between quality degradation due to uncorrected errors and that due to expending bits on error protection (instead of on speech encoding).

Both the GSM and PDC half-rate coders use *convolutional coding* [Proakis, 1995] for error correction. Convolutional codes are sliding or sequential codes. The encoder of a rate $m/n, m < n$ convolutional code can be realized using m shift registers. For every m information bits input to the encoder (one bit to each of the m shift registers), n code bits are output to the channel. Each code bit is computed as a modulo-2 sum of a subset of the bits in the shift registers. Error protection overhead can be reduced by exploiting the unequal sensitivity of speech quality to errors in different positions of the encoded bit stream. A family of *rate-compatible punctured convolutional codes* (RCPCCs) [Hagenauer, 1988] is a collection of related convolutional codes; all of the codes in the collection except the one with the lowest rate are derived by *puncturing* (dropping) code bits from the convolutional code with the lowest rate. With an RCPCC, the channel coding rate can be varied on the fly (i.e., variable-rate coding) while a sequence of information bits is being encoded through the shift registers, thereby imparting on different segments in the sequence different degrees of error protection.

For decoding a convolutional coded bit stream, the *Viterbi algorithm* [Proakis, 1995] is a computationally efficient procedure. Given the output of the demodulator, the algorithm determines the most likely sequence of data bits sent by the channel encoder. To fully utilize the error correction power of the convolutional code, the amplitude of the demodulated *channel symbol* can be quantized to more bits than the minimum number required, i.e., for subsequent *soft decision decoding*. The minimum number of bits is given by the number of channel-coded bits mapped by the modulator onto each channel symbol; decoding based on the minimum-rate bit stream is called *hard decision* decoding. Although soft decoding gives better error protection, decoding complexity is also increased.

Whereas convolutional codes are most effective against randomly scattered bit errors, errors on cellular radio channels often occur in bursts of bits. These bursts can be broken up if the bits put into the channel are rearranged after demodulation. Thus, in *block interleaving*, encoded bits are read into a matrix by row and then read out of the matrix by column (or vice versa) and then passed on to the modulator; the reverse operation is performed by a *deinterleaver* following demodulation. Interleaving increases the transmission delay to the extent that enough bits need to be collected in order to fill up the matrix.

Owing to the severe nature of the cellular radio channel and limited available transmission capacity, uncorrected errors often remain in the decoded data. A common countermeasure is to append an error detection code to the speech data stream prior to channel coding. When residual channel errors are detected, the speech decoder can take various remedial measures to minimize the

Samsung Exhibit 1022, Page 151 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

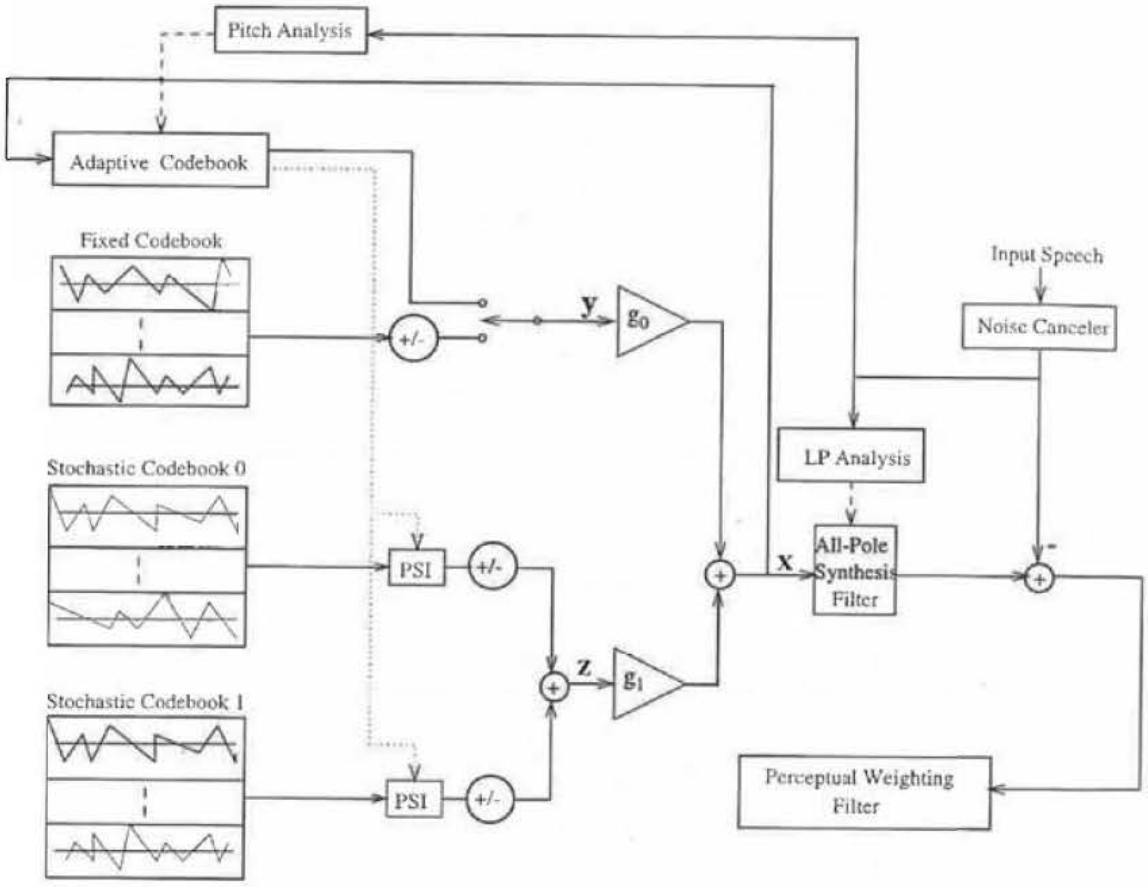


FIGURE 94.2 Basic structure of the PSI-CELP encoder.

negative impact on speech quality. Common measures are repetition of speech parameters from the most recent good frames and gradual muting of the possibly corrupted synthesized speech.

The PDC and GSM half-rate standard algorithms together embody some of the latest advances in speech coding techniques, including: *multimodal coding* where the coder configuration and bit allocation change with the type of speech input; *vector quantization* (*VQ*) [Gersho and Gray, 1991] of the LP filter parameters; higher precision and improved coding efficiency for pitch-periodic excitation; and postfiltering with improved tandeming performance. We next explore the more distinctive features of the PDC and GSM speech coders.

94.6 The Japanese Half-Rate Standard

An algorithm was selected for the Japanese half-rate standard in April 1993, following the evaluation of 12 submissions in a first round, and four final candidates in a second round [Masui and Oguchi, 1993]. The selected algorithm, called pitch synchronous innovation CELP[2] (PSI-CELP), met all of the basic selection criteria and scored the highest among all candidates evaluated. A block diagram of the PSI-CELP encoder is shown in Fig. 94.2, and bit allocations are summarized in Table 94.3. The complexity of the coder is estimated to be approximately 2.4 times that of the PDC full-rate

[2]There were two candidate algorithms named PSI-CELP in the PDC half-rate competition. The algorithm described here was contributed by NTT Mobile Communications Network, Inc. (NTT DoCoMo).

Samsung Exhibit 1022, Page 152 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

96

Wireless LANs

96.1 Introduction 1367
96.2 Physical Layer Design 1368
96.3 MAC Layer Protocols 1369
Reservation-TDMA (R-TDMA) • Distributed Foundation Wireless MAC (DFWMAC) • Randomly Addressed Polling (RAP)
96.4 Network Layer Issues 1373
Alternative View of Mobile Networks • A Proposed Architecture • Networking Issues
96.5 Transport Layer Design 1376
96.6 Conclusions 1378

Suresh Singh
University of South Carolina

96.1 Introduction

A proliferation of high-performance portable computers combined with end-user need for communication is fueling a dramatic growth in wireless **local area network** (LAN) technology. Users expect to have the ability to operate their portable computer globally while remaining connected to communications networks and service providers. Wireless LANs and cellular networks, connected to high-speed networks, are being developed to provide this functionality.

Before delving deeper into issues relating to the design of wireless LANs, it is instructive to consider some scenarios of user mobility.

1. A simple model of user mobility is one where a computer is physically moved while retaining network connectivity at either end. For example, a move from one room to another as in a hospital where the computer is a hand-held device displaying patient charts and the nurse using the computer moves between wards or floors while accessing patient information.
2. Another model situation is where a group of people (at a conference, for instance) set up an ad-hoc LAN to share information as in Fig. 96.1.
3. A more complex model is one where several computers in constant communication are in motion and continue to be networked. For example, consider the problem of having robots in space collaborating to retrieve a satellite.

A great deal of research has focused on dealing with physical layer and **medium access control** (MAC) layer protocols. In this chapter we first summarize standardization efforts in these areas. The remainder of the chapter is then devoted to a discussion of networking issues involved in wireless LAN design. Some of the issues discussed include routing in wireless LANs (i.e., how does data find its destination when the destination is mobile?) and the problem of providing service guarantees

0-8493-8349-4/97/$0.00+$.50

© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 153 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

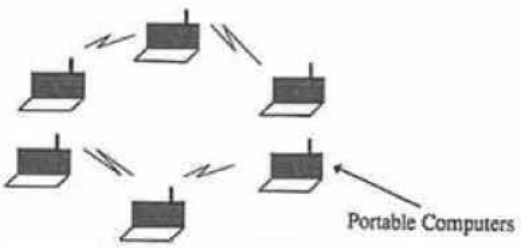


FIGURE 96.1 Ad-hoc wireless LAN.

to end users (e.g., error-free data transmission or bounded delay and bounded bandwidth service, etc.).

96.2 Physical Layer Design

Two media are used for transmission over wireless LANs, infrared and radio frequency. RF LANs are typically implemented in the industrial, scientific, and medical (ISM) frequency bands 902–928 MHz, 2400–2483.5 MHz and 5725–5850 MHz. These frequencies do not require a license allowing the LAN product to be portable, i.e., a LAN can be moved without having to worry about licensing.

IR and RF technologies have different design constraints. IR receiver design is simple (and thus inexpensive) in comparison to RF receiver design because IR receivers only detect the amplitude of the signal not the frequency or phase. Thus, a minimal of filtering is required to reject interference. Unfortunately, however, IR shares the electromagnetic spectrum with the sun and incandescent or fluorescent light. These sources of modulated infrared energy reduce the signal-to-noise ratio of IR signals and, if present in extreme intensity, can make the IR LANs inoperable. There are two approaches to building IR LANs.

1. The transmitted signal can be focused and aimed. In this case the IR system can be used outdoors and has an area of coverage of a few kilometers.
2. The transmitted signal can be bounced off the ceiling or radiated omnidirectionally. In either case, the range of the IR source is 10–20 m (i.e., the size of one medium-sized room).

RF systems face harsher design constraints in comparison to IR systems for several reasons. The increased demand for RF products has resulted in tight regulatory constraints on the allocation and use of allocated bands. In the U.S., for example, it is necessary to implement spectrum spreading for operation in the ISM bands. Another design constraint is the requirement to confine the emitted spectrum to a band, necessitating amplification at higher carrier frequencies, frequency conversion using precision local oscillators, and selective components. RF systems must also cope with environmental noise that is either naturally occurring, for example, atmospheric noise or man made, for example, microwave ovens, copiers, laser printers, or other heavy electrical machinery. RF LANs operating in the ISM frequency ranges also suffer interference from amateur radio operators.

Operating LANs indoors introduces additional problems caused by multipath propagation, Rayleigh fading, and absorption. Many materials used in building construction are opaque to IR radiation resulting in incomplete coverage within rooms (the coverage depends on obstacles within the room that block IR) and almost no coverage outside closed rooms. Some materials, such as white plasterboard, can also cause reflection of IR signals. RF is relatively immune to absorption and reflection problems. Multipath propagation affects both IR and RF signals. The technique to alleviate the effects of multipath propagation in both types of systems is the same use of aimed (directional) systems for transmission enabling the receiver to reject signals based on their angle of incidence. Another technique that may be used in RF systems is to use multiple antennas. The phase difference between different paths can be used to discriminate between them.

Samsung Exhibit 1022, Page 154 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

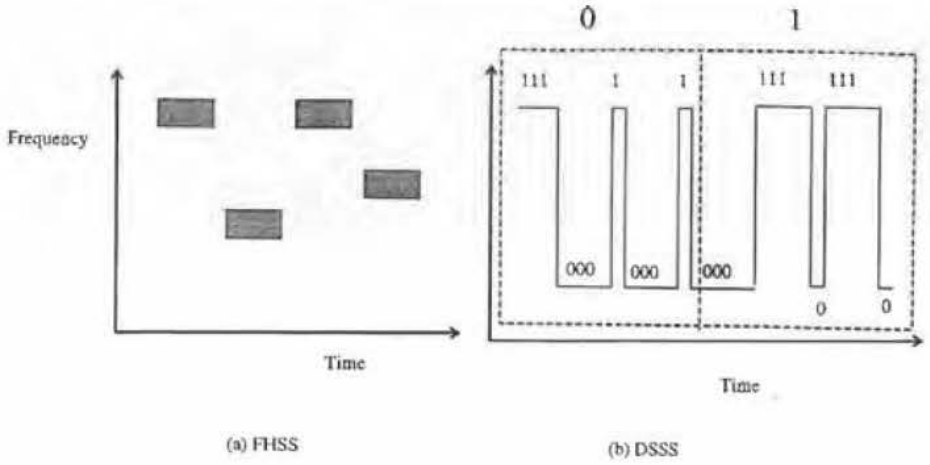


FIGURE 96.2 Spread spectrum.

Rayleigh fading is a problem in RF systems. Recall that Rayleigh fading occurs when the difference in path length of the same signal arriving along different paths is a multiple of half a wavelength. This causes the signal to be almost completely canceled out at the receiver. Because the wavelengths used in IR are so small, the effect of Rayleigh fading is not noticeable in those systems. RF systems, on the other hand, use wavelengths of the order of the dimension of a laptop. Thus, moving the computer a small distance could increase/decrease the fade significantly.

Spread spectrum transmission technology is used for RF-based LANs and it comes in two varieties: direct-sequence spread spectrum (DSSS) and frequency-hopping spread spectrum (FHSS). In a FHSS system, the available band is split into several channels. The transmitter transmits on one channel for a fixed time and then hops to another channel. The receiver is synchronized with the transmitter and hops in the same sequence; see Fig. 96.2(a). In DSSS systems, a random binary string is used to modulate the transmitted signal. The relative rate between this sequence and user data is typically between 10 and 100; see Fig. 96.2(b).

The key requirements of any transmission technology is its robustness to noise. In this respect DSSS and FHSS show some differences. There are two possible sources of interference for wireless LANs: the presence of other wireless LANs in the same geographical area (i.e., in the same building, etc.) and interference due to other users of the ISM frequencies. In the latter case, FHSS systems have a greater ability to avoid interference because the hopping sequence could be designed to prevent potential interference. DSSS systems, on the other hand, do exhibit an ability to recover from interference because of the use of the spreading factor [Fig. 96.2(b)].

It is likely that in many situations several wireless LANs may be collocated. Since all wireless LANs use the same ISM frequencies, there is a potential for a great deal of interference. To avoid interference in FHSS systems, it is necessary to ensure that the hopping sequences are orthogonal. To avoid interference in DSSS systems, on the other hand, it is necessary to allocate different channels to each wireless LAN. The ability to avoid interference in DSSS systems is, thus, more limited in comparison to FHSS systems because FHSS systems use very narrow subchannels (1 MHz) in comparison to DSSS systems that use wider subchannels (for example, 25 MHz), thus, limiting the number of wireless LANs that can be collocated. A summary of design issues can be found in Bantz and Bauchot, 1994.

96.3 MAC Layer Protocols

MAC protocol design for wireless LANs poses new challenges because of the in-building operating environment for these systems. Unlike wired LANs (such as the ethernet or token ring), wireless LANs operate in strong multipath fading channels where channel characteristics can change in very

Samsung Exhibit 1022, Page 155 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

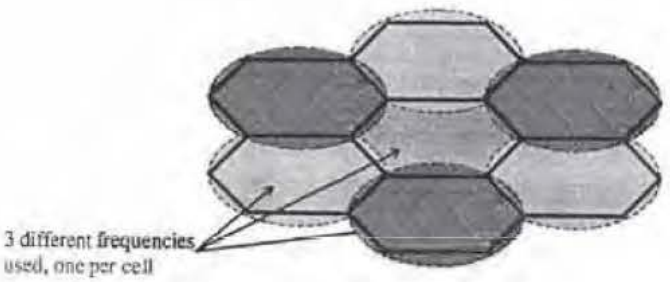


FIGURE 96.3 Cellular structure for wireless LANs (note frequency reuse).

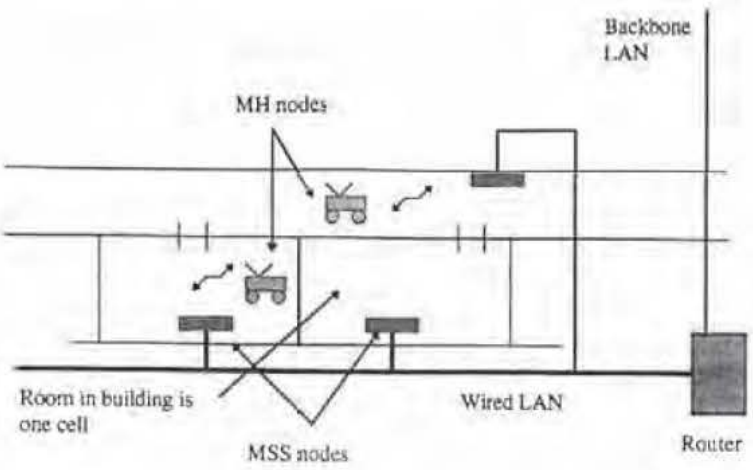


FIGURE 96.4 In-building LAN (made up of several wireless LANs).

short distances resulting in unreliable communication and unfair channel access due to capture. Another feature of the wireless LAN environment is that carrier sensing takes a long time in comparison to wired LANs; it typically takes between 30 and 50 μs (see Chen and Lee, 1994), which is a significant portion of the packet transmission time. This results in inefficiencies if the CSMA family of protocols is used without any modifications.

Other differences arise because of the mobility of users in wireless LAN environments. To provide a building (or any other region) with wireless LAN coverage, the region to be covered is divided into cells as shown in Fig. 96.3. Each cell is one wireless LAN, and adjacent cells use different frequencies to minimize interference. Within each cell there is an access point called a **mobile support station** (MSS) or base station that is connected to some wired network. The mobile users are called **mobile hosts (MH)**. The MSS performs the functions of channel allocation and providing connectivity to existing wired networks; see Fig. 96.4. Two problems arise in this type of an architecture that are not present in wired LANs.

1. The number of nodes within a cell changes dynamically as users move between cells. How can the channel access protocol dynamically adapt to such changes efficiently?
2. When a user moves between cells, the user has to make its presence known to the other nodes in the cell. How can this be done without using up too much bandwidth? The protocol used to solve this problem is called a handoff protocol and works along the following lines: A switching station (or the MSS nodes working together, in concert) collects signal strength information for each mobile host within each cell. Note that if a mobile host is near a

Samsung Exhibit 1022, Page 156 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

cell boundary, the MSS node in its current cell as well as in the neighboring cell can hear its transmissions and determine signal strengths. If the mobile host is currently under the coverage of MSS M1 but its signal strength at MSS M2 becomes larger, the switching station initiates a handoff whereby the MH is considered as part of M2's cell (or network).

The mode of communication in wireless LANs can be broken in two: communication from the mobile to the MSS (called *uplink* communication) and communication in the reverse direction (called *downlink* communication). It is estimated that downlink communication accounts for about 70–80% of the total consumed bandwidth. This is easy to see because most of the time users request files or data in other forms (image data, etc.) that consume much more transmission bandwidth than the requests themselves. In order to make efficient use of bandwidth (and, in addition, guarantee service requirements for real-time data), most researchers have proposed that the downlink channel be controlled entirely by the MSS nodes. These nodes allocate the channel to different mobile users based on their current requirements using a protocol such as **time division multiple access** (TDMA). What about uplink traffic? This is a more complicated problem because the set of users within a cell is dynamic, thus making it infeasible to have a static channel allocation for the uplink. This problem is the main focus of MAC protocol design.

What are some of the design requirements of an appropriate MAC protocol? The IEEE 802.11 recommended standard for wireless LANs has identified almost 20 such requirements, some of which are discussed here (the reader is referred to Chen, 1994, for further details). Clearly any protocol must maximize throughput while minimizing delays and providing fair access to all users. In addition to these requirements, however, mobility introduces several new requirements.

1. The MAC protocol must be independent of the underlying physical layer transmission technology adopted (be it DSSS, FHSS or IR).
2. The maximum number of users can be as high as a few hundred in a wireless LAN. The MAC protocol must be able to handle many users without exhibiting catastrophic degradation of service.
3. The MAC protocols must provide secure transmissions because the wireless medium is easy to tap.
4. The MAC protocol needs to work correctly in the presence of collocated networks.
5. It must have the ability to support ad-hoc networking (as in Fig. 96.1).
6. Other requirements include the need to support priority traffic, preservation of packet order, and an ability to support multicast.

Several contention-based protocols currently exist that could be adapted for use in wireless LANs. The protocols currently being looked by IEEE 802.11 include protocols based on **carrier sense multiple access (CSMA)**, polling, and TDMA. Protocols based on **code division multiple access** (CDMA) and **frequency division multiple access** (FDMA) are not considered because the processing gains obtained using these protocols are minimal while, simultaneously, resulting in a loss of flexibility for wireless LANs.

It is important to highlight an important difference between networking requirements of ad-hoc networks (as in Fig. 96.1) and networks based on cellular structure. In cellular networks, all communication occurs between the mobile hosts and the MSS (or base station) within that cell. Thus, the MSS can allocate channel bandwidth according to requirements of different nodes, i.e., we can use centralized channel scheduling for efficient use of bandwidth. In ad-hoc networks there is no such central scheduler available. Thus, any multiaccess protocol will be contention based with little explicit scheduling. In the remainder of this section we focus on protocols for cell-based wireless LANs only.

All multiaccess protocols for cell-based wireless LANs have a similar structure; see Chen, 1994.

1. The MSS announces (explicitly or implicitly) that nodes with data to send may contend for the channel.

Samsung Exhibit 1022, Page 157 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

2. Nodes interested in sending data contend for the channel using protocols such as CSMA.
3. The MSS allocates the channel to successful nodes.
4. Nodes transmit packets (contention-free transmission).
5. MSS sends an explicit acknowledgment (ACK) for packets received.

Based on this model we present three MAC protocols.

Reservation-TDMA (R-TDMA)

This approach is a combination of TDMA and some contention protocol (see PRMA in Goodman, 1990). The MSS divides the channel into slots (as in TDMA), which are grouped into frames. When a node wants to transmit it needs to reserve a slot that it can use in every consecutive frame as long as it has data to transmit. When it has completed transmission, other nodes with data to transmit may contend for that free slot. There are four steps to the functioning of this protocol.

a. At the end of each frame the MSS transmits a feedback packet that informs nodes of the current reservation of slots (and also which slots are free). This corresponds to steps 1 and 3 from the preceding list.
b. During a frame, all nodes wishing to acquire a slot transmit with a probability ρ during a free slot. If a node is successful it is so informed by the next feedback packet. If more than one node transmits during a free slot, there is a collision and the nodes try again during the next frame. This corresponds to step 2.
c. A node with a reserved slot transmits data during its slot. This is the contention-free transmission (step 4).
d. The MSS sends ACKs for all data packets received correctly. This is step 5.

The R-TDMA protocol exhibits several nice properties. First and foremost, it makes very efficient use of the bandwidth, and average latency is half the frame size. Another big benefit is the ability to implement power conserving measures in the portable computer. Since each node knows when to transmit (nodes transmit during their reserved slot only) it can move into a power-saving mode for a fixed amount of time, thus increasing battery life. This feature is generally not available in CSMA-based protocols. Furthermore, it is easy to implement priorities because of the centralized control of scheduling. One significant drawback of this protocol is that it is expensive to implement (see Barke and Badrinath, 1994).

Distributed Foundation Wireless MAC (DFWMAC)

The CSMA/CD protocol has been used with great success in the ethernet. Unfortunately, the same protocol is not very efficient in a wireless domain because of the problems associated with cell interference (i.e., interference from neighboring cells), the relatively large amount of time taken to sense the channel (see Glisic, 1991) and the hidden terminal problem (see Tobagi and Kleinrock, 1975a and 1975b). The current proposal is based on a CSMA/collision avoidance (CA) protocol with a four-way handshake; see Fig. 96.5.

The basic operation of the protocol is simple. All MH nodes that have packets to transmit compete for the channel by sending ready to transmit (RTS) messages using nonpersistent CSMA. After a station succeeds in transmitting a RTS, the MSS sends a clear to transmit (CTS) to the MH. The MH transmits its data and then receives an ACK. The only possibility of collision that exists is in the RTS phase of the protocol and inefficiencies occur in the protocol, because of the RTS and CTS stages. Note that unlike R-TDMA it is harder to implement power saving functions. Furthermore, latency is dependent on system load making it harder to implement real-time guarantees. Priorities are also not implemented. On the positive side, the hardware for this protocol is very inexpensive.

Samsung Exhibit 1022, Page 158 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

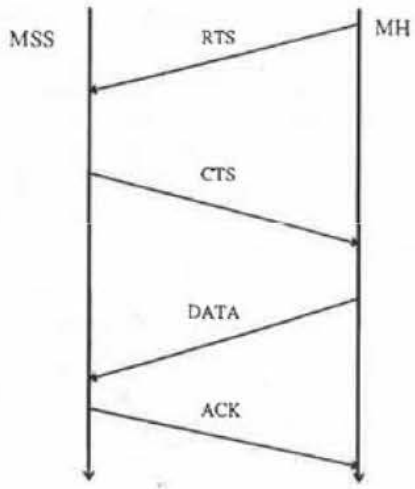


FIGURE 96.5 CSMA/CA and four-way handshaking protocol.

Randomly Addressed Polling (RAP)

In this scheme, when a MSS is ready to collect uplink packets it transmits a READY message. At this point all nodes with packets to send attempt to grab the channel as follows.

a. Each MH with a packet to transmit generates a random number between 0 and P.
b. All active MH nodes simultaneously and orthogonally transmit their random numbers (using CDMA or FDMA). We assume that all of these numbers are received correctly by the MSS. Remember that more than one MH node may have selected the same random number.
c. Steps a and b are repeated L times.
d. At the end of L stages, the MSS determines a stage (say, k) where the total number of distinct random numbers was the largest. The MSS polls each distinct each random number in this stage in increasing order. All nodes that had generated the polled random number transmit packets to the MSS.
e. Since more than one node may have generated the same random number, collisions are possible. The MSS sends a ACK or NACK after each such transmission. Unsuccessful nodes try again during the next iteration of the protocol.

The protocol is discussed in detail in Chen and Lee, 1993 and a modified protocol called GRAP (for group RAP) is discussed in Chen, 1994. The authors propose that GRAP can also be used in the contention stage (step 2) for TDMA- and CSMA-based protocols.

96.4 Network Layer Issues

An important goal of wireless LANs is to allow users to move about freely while still maintaining all of their connections (network resources permitting). This means that the network must route all packets destined for the mobile user to the MSS of its current cell in a transparent manner. Two issues need to be addressed in this context.

- How can users be addressed?
- How can active connections for these mobile users be maintained?

Ioanidis, Duchamp, and Maguire, 1991 propose a solution called the IPIP (IP-within-IP) protocol. Here each MH has a unique **internet protocol** (IP) address called its home address. To deliver

Samsung Exhibit 1022, Page 159 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

a packet to a remote MH, the source MSS first broadcasts an address resolution protocol (ARP) request to all other MSS nodes to locate the MH. Eventually some MSS responds. The source MSS then encapsulates each packet from the source MH within another packet containing the IP address of the MSS in whose cell the MH is located. The destination MSS extracts the packet and delivers it to the MH. If the MH has moved away in the interim, the new MSS locates the new location of the MH and performs the same operation. This approach suffers from several problems as discussed in Teraoka and Tokoro, 1993. Specifically, the method is not scaleable to a network spanning areas larger than a campus for the following reasons.

1. IP addresses have a prefix identifying the campus subnetwork where the node lives; when the MH moves out of the campus, its IP address no longer represents this information.
2. The MSS nodes serve the function of routers in the mobile network and, therefore, have the responsibility of tracking all of the MH nodes globally causing a lot of overhead in terms of message passing and packet forwarding; see Ghai and Singh, 1994.

Teraoka and Tokoro, 1993, have proposed a much more flexible solution to the problem called virtual IP (VIP). Here every mobile host has a virtual IP address that is unchanging regardless of the location of the MH. In addition, hosts have physical network addresses (traditional IP addresses) that may change as the host moves about. At the transport layer, the target node is always specified by its VIP address only. The address resolution from the VIP address to the current IP address takes place either at the network layer of the same machine or at a gateway. Both the host machines and the gateways maintain a cache of VIP to IP mappings with associated timestamps. This information is in the form of a table called *address mapping table* (AMT). Every MH has an associated *home gateway*. When a MH moves into a new subnetwork, it is assigned a new IP address. It sends this new IP address and its VIP address to its home gateway via a *VipConn* control message. All intermediate gateways that relay this message update their AMT tables as well. During this process of updating the AMT tables, all packets destined to the MH continue to be sent to the old location. These packets are returned to the sender, who then sends them to the home gateway of the MH. It is easy to see that this approach is easily scaleable to large networks, unlike the IPIP approach.

Alternative View of Mobile Networks

The approaches just described are based on the belief that mobile networks are merely an extension of wired networks. Other authors [Singh, 1995] disagree with this assumption because there are fundamental differences between the mobile domain and the fixed wired network domain. Two examples follow.

1. The available bandwidth at the wireless link is small; thus, end-to-end packet retransmission for transmission control protocol (TCP)-like protocols (implemented over datagram networks) is a bad idea. This leads to the conclusion that transmission within the mobile network must be connection oriented. Such a solution, using virtual circuits (VC), is proposed in Ghai and Singh, 1994.
2. The bandwidth available for a MH with open connections changes dynamically since the number of other users present in each cell varies randomly. This is a feature not present in fixed high-speed networks where, once a connection is set up, its bandwidth does not vary much. Since bandwidth changes are an artifact of mobility and are dynamic, it is necessary to deal with the consequences (e.g., buffer overflow, large delays, etc.) locally to both, i.e., shield fixed network hosts from the idiosyncrasies of mobility as well as to respond to changing bandwidth quickly (without having to rely on end-to-end control). Some other differences are discussed in Singh, 1995.

Samsung Exhibit 1022, Page 160 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

A Proposed Architecture

Keeping these issues in mind, a more appropriate architecture has been proposed in Ghai and Singh, 1994, and Singh, 1995. Mobile networks are considered to be different and separate from wired networks. Within a mobile network is a three-layer hierarchy; see Fig. 96.6. At the bottom layer are the MHs. At the next level are the MSS nodes (one per cell). Finally, several MSS nodes are controlled by a **supervisor host (SH)** node (there may be one SH node per small building). The SH nodes are responsible for flow control for all MH connections within their domain; they are also responsible for tracking MH nodes and forwarding packets as MH nodes roam. In addition, the SH nodes serve as a *gateway* to the wired networks. Thus, any connection setup from a MH to a fixed host is broken in two, one from the MH to the SH and another from the SH to the fixed host. The MSS nodes in this design are simply connection endpoints for MH nodes. Thus, they are simple devices that implement the MAC protocols and little else. Some of the benefits of this design are as follows.

1. Because of the large coverage of the SH (i.e., a SH controls many cells) the MH remains in the domain of one SH much longer. This makes it easy to handle the consequences of dynamic bandwidth changes locally. For instance, when a MH moves into a crowded cell, the bandwidth available to it is reduced. If it had an open ftp connection, the SH simply buffers undelivered packets until they can be delivered. There is no need to inform the other endpoint of this connection of the reduced bandwidth.
2. When a MH node sets up a connection with a service provider in the fixed network, it negotiates some quality of service (QOS) parameters such as bandwidth, delay bounds, etc. When the MH roams into a crowded cell, these QOS parameters can no longer be met because the available bandwidth is smaller. If the traditional view is adopted (i.e., the mobile networks are extensions of fixed networks) then these QOS parameters will have to be *renegotiated* each time the bandwidth changes (due to roaming). This is a very expensive proposition because of the large number of control messages that will have to be exchanged. In the approach of Singh, 1995, the service provider will never know about the bandwidth changes since it deals only with the SH that is accessed via the wired network. The SH bears the responsibility of handling bandwidth changes by either buffering packets until the bandwidth available to the MH increases (as in the case of the ftp example) or it could discard a fraction of real-time packets (e.g., a voice connection) to ensure delivery of most of the packets within their deadlines. The SH could also instruct the MSS to allocate a larger amount of bandwidth to the MH when the number of buffered packets becomes large. Thus, the service provider in the fixed network is shielded from the mobility of the user.

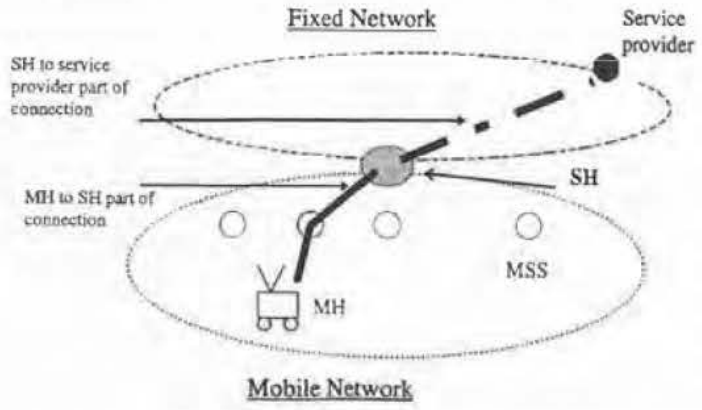


FIGURE 96.6 Proposed architecture for wireless networks.

Samsung Exhibit 1022, Page 161 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Networking Issues

It is important for the network to provide connection-oriented service in the mobile environment (as opposed to connectionless service as in the internet) because bandwidth is at a premium in wireless networks, and it is, therefore, inadvisable to have end-to-end retransmission of packets (as in TCP). The proposed architecture is well suited to providing connection-oriented service by using VCs.

In the remainder of this section we look at how virtual circuits are used within the mobile network and how routing is performed for connections to mobile hosts. Every connection set up with one or more MH nodes as a connection endpoint is routed through the SH nodes and each connection is given a unique VC number. The SH node keeps track of all MH nodes that lie within its domain. When a packet needs to be delivered to a MH node, the SH first buffers the packet and then sends it to the MSS at the current location of the MH or to the predicted location if the MH is currently between cells. The MSS buffers all of these packets for the MH and transmits them to the MH if it is in its cell. The MSS discards packets after transmission or if the SH asks it to discard the packets. Packets are delivered in the correct order to the MH (without duplicates) by having the MH transmit the expected sequence number (for each VC) during the initial handshake (i.e., when the MH first enters the cell). The MH sends ACKs to the SH for packets received. The SH discards all packets that have been acknowledged. When a MH moves from the domain of SH1 into the domain of SH2 while having open connections, SH1 continues to forward packets to SH2 until either the connections are closed or until SH2 sets up its own connections with the other endpoints for each of MH's open connections (it also gives new identifiers to all these open connections). The detailed protocol is presented in Ghai and Singh, 1994.

The SH nodes are all connected over the fixed (wired) network. Therefore, it is necessary to route packets between SH nodes using the protocol provided over the fixed networks. The VIP protocol appears to be best suited to this purpose. Let us assume that every MH has a globally unique VIP address. The SHs have both a VIP as well as a fixed IP address. When a MH moves into the domain of a SH, the IP address affixed to this MH is the IP address of the SH. This ensures that all packets sent to the MH are routed through the correct SH node. The SH keeps a list of all VIP addresses of MH nodes within its domain and a list of open VCs for each MH. It uses this information to route the arriving packets along the appropriate VC to the MH.

96.5 Transport Layer Design

The transport layer provides services to higher layers (including the application layer), which include connectionless services like UDP or connection-oriented services like TCP. A wide variety of new services will be made available in the high-speed networks, such as continuous media service for real-time data applications such as voice and video. These services will provide bounds on delay and loss while guaranteeing some minimum bandwidth.

Recently variations of the TCP protocol have been proposed that work well in the wireless domain. These proposals are based on the traditional view that wireless networks are merely extensions of fixed networks. One such proposal is called I-TCP [Barke and Badrinath, 1994] for indirect TCP. The motivation behind this work stems from the following observation. In TCP the sender times out and begins retransmission after a timeout period of several hundred milliseconds. If the other endpoint of the connection is a mobile host, it is possible that the MH is disconnected for a period of several seconds (while it moves between cells and performs the initial greeting). This results in the TCP sender timing out and transmitting the same data several times over, causing the effective throughput of the connection to degrade rapidly. To alleviate this problem, the implementation of I-TCP separates a TCP connection into two pieces—one from the fixed host to another fixed host that is near the MH and another from this host to the MH (note the similarity of this approach with the approach in Fig. 96.6). The host closer to the MH is aware of mobility and has a larger timeout

Samsung Exhibit 1022, Page 162 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

period. It serves as a type of gateway for the TCP connection because it sends ACKs back to the sender before receiving ACKs from the MH. The performance of I-TCP is far superior to traditional TCP for the mobile networks studied.

In the architecture proposed in Fig. 96.6, a TCP connection from a fixed host to a mobile host would terminate at the SH. The SH would set up another connection to the MH and would have the responsibility of transmitting all packets correctly. In a sense this is a similar idea to I-TCP except that in the wireless network VCs are used rather than datagrams. Therefore, the implementation of TCP service is made much easier.

A problem that is unique to the mobile domain occurs because of the unpredictable movement of MH nodes (i.e., a MH may roam between cells resulting in a large variation of available bandwidth in each cell). Consider the following example. Say nine MH nodes have opened 11-kb/s connections in a cell where the available bandwidth is 100 kb/s. Let us say that a tenth mobile host M10, also with an open 11-kb/s connection, wanders in. The total requested bandwidth is now 110 kb/s while the available bandwidth is only 100 kb/s. What is to be done? One approach would be to deny service to M10. However, this seems an unfair policy. A different approach is to penalize all connections equally so that each connection has 10-kb/s bandwidth allocated.

To reduce the bandwidth for each connection from 11 kb/s to 10 kb/s, two approaches may be adopted:

1. Throttle back the sender for each connection by sending control messages.
2. Discard 1-kb/s data for each connection at the SH. This approach is only feasible for applications that are tolerant of data loss (e.g., real-time video or audio).

The first approach encounters a high overhead in terms of control messages and requires the sender to be capable of changing the data rate dynamically. This may not always be possible; for instance, consider a teleconference consisting of several participants where each mobile participant is subject to dynamically changing bandwidth. In order to implement this approach, the data (video or audio or both) will have to be compressed at different ratios for each participant, and this compression ratio may have to be changed dynamically as each participant roams. This is clearly an unreasonable solution to the problem. The second approach requires the SH to discard 1-kb/s of data for each connection. The question is, how should this data be discarded? That is, should the 1 kb of discarded data be consecutive (or clustered) or uniformly spread out over the data stream every 1 s? The way in which the data is discarded has an effect on the final perception of the service by the mobile user. If the service is audio, for example, a random uniform loss is preferred to a clustered loss (where several consecutive words are lost). If the data is compressed video, the problem is even more serious because most random losses will cause the encoded stream to become unreadable resulting in almost a 100% loss of video at the user.

A solution to this problem is proposed in Seal and Singh, 1995, where a new sublayer is added to the transport layer called the *Loss profile transport sublayer* (*LPTSL*). This layer determines how data is to be discarded based on special transport layer markers put by application calls at the sender and based on negotiated loss functions that are part of the QOS negotiations between the SH and service provider. Figure 96.7 illustrates the functioning of this layer at the service provider, the SH, and the MH. The original data stream is broken into *logical segments* that are separated by markers (or flags). When this stream arrives at the SH, the SH discards entire logical segments (in the case of compressed video, one logical segment may represent one frame) depending on the bandwidth available to the MH. The purpose of discarding entire logical segments is that discarding a part of such a segment of data makes the rest of the data within that segment useless—so we might as well discard the entire segment. Observe also that the flags (to identify logical segments) are inserted by the LPTSL via calls made by the application layer. Thus, the transport layer or the LPTSL does not need to know encoding details of the data stream. This scheme is currently being implemented at the University of South Carolina by the author and his research group.

Samsung Exhibit 1022, Page 163 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

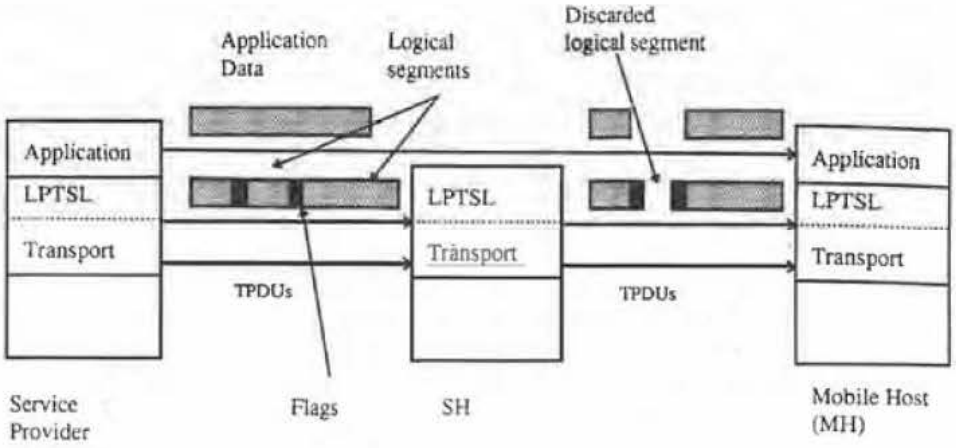


FIGURE 96.7 LPTSL, an approach to handle dynamic bandwidth variations.

96.6 Conclusions

The need for wireless LANs is driving rapid development in this area. The IEEE has proposed standards (802.11) for the physical layer and MAC layer protocols. A great deal of work, however, remains to be done at the network and transport layers. There does not appear to be a consensus regarding subnet design for wireless LANs. Our work has indicated a need for treating wireless LAN subnetworks as being fundamentally different from fixed networks, thus resulting in a different subnetwork and transport layer designs. Current efforts are underway to validate these claims.

Defining Terms

Carrier-sense multiple access (CSMA): Protocols such as those used over the ethernet.
Medium access control (MAC): Protocols arbitrate channel access between all nodes on a wireless LAN.
Mobile host (MH) nodes: The nodes of wireless LAN.
Supervisor host (SH): The node that takes care of flow-control and other protocol processing for all connections.

References

Bantz, D.F. and Bauchot, F.J. 1994. Wireless LAN design alternatives. *IEEE Network.* 8(2):43–53.
Barke, A. and Badrinath, B.R. 1994. I-TCP: indirect TCP for mobile hosts. Tech. Rept. DCS-TR-314, Dept. Computer Science, Rutgers Univ. Piscataway, NJ.
Chen, K.-C. 1994. Medium access control of wireless LANs for mobile computing. *IEEE Network,* 8(5):50–63.
Chen, K.-C. and Lee, C.H. 1993. RAP: a novel medium access control protocol for wireless data networks. *Proc. IEEE GLOBECOM'93,* IEEE Press, Piscataway, NJ 08854. 1713–1717.
Ghai, R. and Singh, S. 1994. An architecture and communication protocol for picocellular networks. *IEEE Personal Comm. Mag.* 1(3):36–46.
Glisic, S.G. 1991. 1-Persistent carrier sense multiple access in radio channel with imperfect carrier sensing. *IEEE Trans. on Comm.* 39(3):458–464.
Goodman, D.J. 1990. Cellular packet communications. *IEEE Trans. on Comm.* 38(8):1272–1280.
Ioanidis, J., Duchamp, D., and Maguire, G.Q. 1991. IP-based protocols for mobile internetworking. *Proc. of ACM SIGCOMM'91,* ACM Press, New York, NY 10036 (Sept.):235–245.

Samsung Exhibit 1022, Page 164 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Seal, K. and Singh, S. 1995. Loss profiles: a quality of service measure in mobile computing. *J. Wireless Networks*, submitted for publication.

Singh, S. 1995. Quality of service guarantees in mobile computing. *J. of Computer Comm.*, to appear.

Teraoka, F. and Tokoro, M. 1993. Host migration transparency in IP networks: the VIP approach. *Proc. of ACM SIGCOMM*, ACM Press, New York, NY 10036 (Jan.):45–65.

Tobagi, F. and Kleinrock, L. 1975a. Packet switching in radio channels: Part I carrier sense multiple access models and their throughput delay characteristic. *IEEE Trans. on Comm.*, 23(12):1400–1416.

Tobagi, F. and Kleinrock, L. 1975b. Packet switching in radio channels: Part II the hidden terminal problem in CSMA and busy-one solution. *IEEE Trans. on Comm.*, 23(12):1417–1433.

urther Information

A good introduction to physical layer issues is presented in Bantz, 1994 and MAC layer issues are discussed in Chen, 1994. For a discussion of network and transport layer issues, see Singh, 1995 and Ghai and Singh, 1994.

Samsung Exhibit 1022, Page 165 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

imagery, neither general purpose stochastic models of the source nor mathematical fidelity criteria that capture all of the known properties of source and receiver are known at this time. Thus, optimal source coding for time-varying imagery does not currently rest on a firm theoretical basis. Rather, we have available an arsenal of tools that work well and that address, at least qualitatively, the known properties of the source and receiver.

The purpose of this chapter is to identify the properties of source and receiver that are exploited by video source coding algorithms, and to describe the tools that are available for source coding. The emphasis is on those aspects that are specific to time-varying imagery, since a previous section deals with still imagery. Thus, the use of motion information in the processing and compression of time-varying imagery is a pivotal element for this contribution. An overview of video compression standards is also given, followed by some perspectives for the future. An extensive treatment of video coding can be found in Netravali and Haskell [1995].

102.2 Source Characteristics and Viewer Requirements

A number of factors determine the source characteristics and the level of redundancy that can be exploited. These include the scanning format, the signal-to-noise ratio, the amount of detail, and the characteristics of the motion. Note that the source signal to be coded may not necessarily be the direct output of a video camera; it may be the result of some scan conversion or other preprocessing operations.

The scanning format encompasses the picture size and the sampling structure. We assume that the input to the source coder is fully digital, that is an ordered sequence of fixed length binary codewords representing the image color at a discrete set of points Ψ. This is denoted

$$u_i(x, y, t), \quad (x, y, t) \in \Psi_i, \qquad i = 1, 2, 3 \tag{102.1}$$

where u_1, u_2, u_3 represent coordinates in a suitable color space. These could be red-green-blue or luminance-chrominance components. Clearly black and white images would be represented by a single component. The sampling structure Ψ is a regular array of points such as a lattice, cropped spatially to the image frame size. Picture size may range from small windows on a videophone or workstation, say, of size 64 by 64, to HDTV pictures of size 1920 by 1080 and beyond. Two vertical-temporal scanning structures are widely used for video: **progressive scanning** and **interlaced scanning**. These scanning structures are shown in vertical-temporal profile in Fig. 102.1. In analog systems, where the capture, transmission, and display formats are the same, interlace is a method to achieve higher vertical resolution for a given vertical scan rate and system bandwidth. The vertical

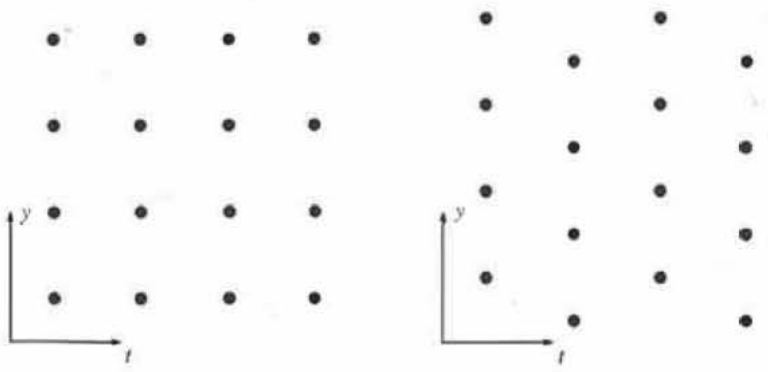


FIGURE 102.1 Profile view of scanning structures: (a) progressive and (b) interlace.

Samsung Exhibit 1022, Page 166 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

TABLE 102.1 Typical Scanning Formats Used for Digital Video Communications

Identifier	Luminance Size		Chrominance Size		Scanning Structure	
	Hor.	Vert.	Hor.	Vert.	Interlace	Pict. Rate
QCIF	176	144	88	72	1:1	29.97
CIF	352	288	176	144	1:1	29.97
ITU-R 601	720	480	360	480	2:1	59.94
HDTV-P	1280	720	640	360	1:1	60.0
HDTV-I	1920	1080	960	540	2:1	60.0

scan rate of 50 or 60 Hz used in major worldwide systems is largely determined by the requirement to minimize the visibility of flicker at the display. Although interlace is a good technique in the context of analog systems, it usually leads to spatiotemporal **aliasing** in the presence of vertical motion, which makes subsequent processing, and especially source coding, much more difficult. (See Dubois [1992] for a discussion.) For this reason, there has been increasing focus on progressive scanning in recent system proposals. A detailed discussion of sampling structures and video signal sampling is given in Dubois [1985].

Although some standards allow considerable flexibility in the choice of picture format, a number of specific formats are in common use or have been proposed. Table 102.1 shows the parameters of some of these formats. The final coded bit rate is very directly related to the source sampling rate, although not linearly.

The signal-to-noise ratio of the source has a significant impact on compression efficiency. In general, noise has little redundancy and cannot be compressed very much; thus, it should be removed prior to coding if possible. There is, of course, a tradeoff between coding efficiency and the introduction of artifacts due to noise reduction. Valid assumptions must be made, or the processing may remove desired picture content such as snowstorms, fireworks, etc. Motion compensated processing is required for the most effective noise reduction [Dubois, 1992]. However, good image and noise models are necessary. The principle is to filter out those deviations from an assumed image model (e.g., that image intensity is constant along trajectories of motion) that match the characteristics of an assumed noise model (e.g., white noise with a specified variance).

The characteristics of the motion in the scene affect the amount of compression that is possible. Apparent motion in the image can be due to either camera motion (including zooming), object motion, or both. In a videophone application, most apparent motion may be due to movement of the person in front of the camera. In a surveillance application, most of it may be due to the scanning motion of the camera (except during an incident), whereas in a sports scene both camera motion and scene motion may be significant. Motion may in general be low, as in a videophone application, or very active, as in a sports application; it may be highly structured (rigid-body motion) or pseudorandom (rustling leaves). Knowledge of the motion characteristics can affect the coding algorithm; if there is unrestricted motion, a worst-case assumption must be made.

Redundancy in a source is related to the concept of predictability (although they are not the same). Given that a certain portion of the source signal has already been transmitted, the coder may attempt to predict what comes next. If the prediction is at all reliable, it is better to transmit information concerning the deviation from the prediction than the raw information. Predictability implies a certain regularity in the data that can be captured in statistical models. In the spatial dimension, if a certain pattern appears over a certain area, one may assume that the pattern will continue, and make the prediction on this basis. Of course, at an object boundary, the prediction will be wrong. However, if it is right most of the time, an overall coding gain can be achieved. In a moving picture, the assumption is that if an object is moving with a certain velocity and acceleration, it will continue to do so. Thus, if the velocity and acceleration are known, the coder can predict what the next frame will look like, and only transmit the error between this prediction and reality. This is the essence of motion-compensated coding. Again, at a scene change, the prediction will

Samsung Exhibit 1022, Page 167 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

be completely wrong, and more information must be transmitted. However, on average there will be a substantial compression. Temporal change is not due solely to motion. The illumination may change, and the luminance of scene points may change as the angles between object surface, illumination source, and camera change. If we can model these effects as well, we can make better predictions and thereby improve coding performance.

In addition to the structural or statistical redundancy in the source, there is perceptual redundancy (also called irrelevancy). Specifically, information that cannot be perceived by the viewer need not be transmitted. The main phenomena related to this are spatial and temporal **masking**. Spatial masking (discussed in the chapter on still image coding) is the effect whereby certain types of distortion are less visible in the vicinity of high-contrast image detail than in flat low-contrast areas. Temporal masking is the effect where distortions are less visible in areas of the image that are newly exposed or contain fast motion. An important case here is the scene change. The viewer can tolerate a significant reduction in resolution just after a scene change. This is fortuitous, because this is exactly where the temporal motion-compensated prediction breaks down.

The goal of an efficient source coder is to maximize quality (or minimize distortion) subject to a number of constraints. In most applications of video compression, the quality or distortion are subjective measures determined by the visual properties of the viewer. However, if further signal processing or manipulation of the image sequence is to be performed, a more stringent objective criterion may be more appropriate, since impairments invisible to a viewer could have a negative effect on a signal processing operation such as chroma key. If subjective quality is the main concern, it would be very desirable to have a mathematical function that measures the subjective distortion. This distortion function could then be minimized during the optimization of the coding algorithm. Unfortunately, such a general purpose and widely accepted perceptual distortion measure is not available. Thus, properties of the human visual system are usually exploited based on qualitative or empirical considerations that are obtained under restricted conditions. However, this approach may still be quite effective [Schreiber, 1993; Jayant, Johnson, and Safranck, 1993]. The coding algorithm can then be evaluated using formal subjective testing under controlled conditions [Allnutt, 1983].

Other requirements may affect the type of coding algorithm that can be used. These include delay, random access, playback control, and scalability. Both the coding algorithm and the transmission channel introduce delay. Depending on the application, there may be tight bounds on this delay, especially in interactive applications. The most stringent application is in face-to-face communication, where a maximum end-to-end delay of about 200 ms can be tolerated. If the delay is longer than this, spontaneous interaction becomes difficult. In a database access application, somewhat longer delays of up to a few seconds may be acceptable. Finally, in an application such as video on demand, the delay constraint may be of no practical consequence, that is, any reasonable encoding system would meet the delay constraint.

Random access and playback control are requirements that may exist in services such as database access and video on demand. Some coding schemes are recursive in nature, and require that the sequence be decoded continuously from the beginning in order to view a particular frame. This may be acceptable in an application such as video conferencing, but it precludes random access and channel switching. Thus, the encoding algorithm must have *finite memory* in some sense to allow such a random access feature. The playback controls include fast-forward, reverse, and fast-reverse playback as well as pause functions. Again, general coding algorithms may not permit such features, and they must be incorporated as constraints if they are desired.

The coding algorithm must account for the properties of the channel or storage medium. The channel may allow only constant bit rate (CBR) operation, or it may permit variable bit rate (VBR) operation. In the former case, the coder must be designed to produce a fixed number of bits over a specified unit of time. In the latter case, the bit rate may vary over time, although some constraints negotiated at setup time must be respected (e.g., ratio of peak to average bit rate). Another requirement of a coding algorithm is acceptable performance in the presence of typical channel impairments. These channel impairments may vary considerably for different types of

Samsung Exhibit 1022, Page 168 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

channels. For example, a dedicated digital link may have random errors with a very low probability of occurrence, whereas a wireless link may have bursty errors with a high probability of occurrence. Also, depending on the transport mechanism, the errors may occur in isolated bits, or they may involve an entire packet, consisting of a large number of bits. In general, when the coding efficiency for redundancy reduction is higher, the coded data is more vulnerable to transmission errors. Thus, overall system performance must be evaluated with the channel impairments taken into account in order to be realistic. If possible, joint source/channel coding should be considered.

In some situations, the same video signal must be delivered to a variety of receivers having different capabilities. Some examples include the broadcast of entertainment television to a population of receivers ranging from small tabletop units to large high-definition displays. Similarly, a database containing video data might be accessed by terminals ranging from relatively inexpensive home computers to powerful workstations, perhaps with three-dimensional display. The capability to transmit only that portion of the data required by the terminal, or to extract a subset of the data, is known as **scalability**. The most straightforward approach to this is simulcasting or multiple storage of the coded data for each type of receiver. This is complex and wasteful of storage or transmission capacity. Thus a more efficient approach is the concept of embedded coding, whereby a subset of the encoded data can be used to decode a lower version of the sequence suitable for display on the lower capability display.

Most of the preceding requirements can reduce the coding efficiency achievable, as compared to an unconstrained situation. In each case, the system designer must decide whether the benefits of the feature justify the additional complexity or performance loss.

02.3 Coding Algorithms

The goal of a source coding algorithm is usually to provide the best image quality for a given channel or data rate, subject to constraints on complexity, delay, or other requirements as discussed. In some circumstances, a certain quality may be prescribed, and a coding algorithm is sought that provides this quality at the lowest possible rate (again, subject to constraints). In time-varying imagery, most incremental information over time is carried by the motion occurring in the scene. Thus, the most highly efficient coding algorithms are based on motion estimation and compensation. This is the main feature distinguishing video coding from still image coding, and it represents the focus of this section. An image sequence is usually encoded by both intraframe coding techniques and interframe coding techniques. The intraframe coding algorithms are very similar to the still picture coding algorithms of the previous section. They are used for several reasons, including: to permit random access into the sequence, to allow fast forward and reverse playback, to provide robustness to channel errors, and as fallback when temporal coding fails (for example, at a scene change).

Motion Compensation

The three-dimensional motion of objects in a scene induces two-dimensional motion in the image plane. The relationship between them depends on the perspective transformation of the camera, and may be complex. The image of a given scene point traces out over time a trajectory in the space–time coordinates of the time-varying image (Fig. 102.2). The basic principle of motion-compensated coding is that the redundancy is very high along this trajectory. This is exploited using the coder structure of Fig. 102.3. The input signal is applied to a motion estimation module that determines a **motion field**. This motion field establishes the correspondences between image samples lying on the same trajectory in successive image pictures. Although the motion field specifies estimated displacements for each pixel in the image sequence, it may be specified parametrically over regions, as will be described subsequently. The motion field is then encoded with a lossless code for transmission to the receiver at rate R_M. Any quantization of the motion field is assumed to take place within the motion estimation module. The motion field is also supplied to the

Samsung Exhibit 1022, Page 169 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

103

The High-Definition Television Grand Alliance System

103.1 Introduction 1462
103.2 System Requirements for Terrestrial Broadcast in the U.S. 1463
103.3 Grand Alliance System Overview 1463
103.4 HDTV System Descriptions 1464
Scanning Formats • Motion Compensation • Motion Estimation • Predictive Motion Compensation Loop • Decoder Loop • Refreshing • Adaptive Selection and Quantization of Coefficients • Forward Analyzer • Variable Length Coding • Buffer Control • Compressed Video Formats • GA System Layer • GA Transmission System
103.5 Scanning Format Applications 1471
HDTV Interoperability with Computers
Migration to 1080 Progressive with 60 Frames Per Second • Compressed Video Storage and Manipulation
103.6 Conclusions 1473

Eric Petajan
Bell Laboratories

103.1 Introduction

In 1987, the FCC chartered an advisory committee to recommend an advanced television system for the U.S. Many proposals were submitted and subsequently withdrawn from consideration or rejected by the Advisory Committee on Advanced Television Service (ACATS). From 1990 to 1992, the Advanced Television Test Center (ATTC) tested four all-digital systems, one analog high-definition television (HDTV) system, and one analog National Television Systems Committee (NTSC) enhancement system. The formation of the Grand Alliance (GA) resulted from the withdrawal by NHK of the only analog HDTV system from the competition and a stalemate between the other four all-digital systems. The members of the GA are AT&T, General Instrument Corporation, Massachusetts Institute of Technology (MIT), Philips Electronics North America Corporation, David Sarnoff Research Center, Thomson Consumer Electronics, and Zenith Electronics Corporation.

Digicipher™ [GI, 1991] and channel compatible Digicipher™ (CCDC) [MIT, 1992] HDTV systems were developed by a consortium between the MIT and General Instrument (GI); the digital spectrum compatible (DSC) HDTV system was developed by Zenith and AT&T; and the advanced digital television (AD-HDTV) [ATRC, 1992] HDTV system was developed by a consortium with Philips, Thomson, Sarnoff, NBC, and Compression Labs Inc. (CLI). The video coder in the

0-8493-8349-8/97/$0.00+$.50
© 1997 by CRC Press, Inc.

Samsung Exhibit 1022, Page 170 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

AD-HDTV system was based on the Moving Pictures Expert Group-1 (MPEG-1) [ISO, 1991; Le Gall, 1991] standard. The GA system [ATSC, 1995] is based on the MPEG-2 [ISO/IEC, 1994a, 1994b] standard and also contains features from all of the original systems. These features are described with particular emphasis on the video coding system.

103.2 System Requirements for Terrestrial Broadcast in the U.S.

The Federal Communications Commission (FCC) has declared that 6-MHz **terrestrial broadcast** channels will be allocated for HDTV[Service, 1993]. Most of these channels are NTSC **taboo channels** which contain interference from distant NTSC transmitters in the same band or adjacent channel interference from nearby transmitters. The FCC has required that the selected HDTV system provide acceptable viewer coverage without causing excessive interference with frequency-collocated or adjacent NTSC channels.

The selected HDTV system should provide greatly improved picture quality compared to NTSC. The spatial resolution should be at least twice that of NTSC horizontally (H) and vertically (V) without exhibiting **interlace artifacts**, and the HDTV system standard must avoid the **chrominance** artifacts and poor chrominance fidelity associated with NTSC. In addition, the audio must be delivered digitally using multichannel compression. Finally, HDTV will have a 16×9 aspect ratio for compatibility with film (compared to 4×3 for NTSC) . These bandwidth and picture quality requirements demand the use of video compression techniques with the highest possible performance.

Receiver complexity should be minimized to ensure that manufacturers can eventually provide consumers with receivers at a cost that will not inhibit the proliferation of HDTV. This constraint implies that video coding and transmission techniques that place most of the complexity in the encoder rather than the decoder are preferred.

103.3 Grand Alliance System Overview

The GA system contains algorithms and components from each of the original digital systems. For example, perceptual modeling, bidirectional motion compensation, hierarchical motion estimation, and adaptive preprocessing, shown in Fig. 103.1, were used in one or more or the first systems. In addition, the Dolby AC-3 [ATSC, 1995] audio compression system was adopted by the GA and was previously used by General Instrument and Zenith/AT&T. The transmission system uses vestigial sideband modulation (VSB) with trellis coding and Reed–Solomon coding for error detection/correction. Testing of the prototype system is complete and the ACATS has issued its recommendation in favor of the Grand Alliance proposal. A U.S. HDTV standard is anticipated sometime in 1996.

One of the biggest challenges before the GA is to provide high picture quality given the bit rate constraints of terrestrial broadcasting in a 6-MHz channel. The presence of significant levels of noise and NTSC interference, and FCC service requirements limits the bit rate to about 20 Mb/s. Two of the key features of the GA system that help meet this challenge are the adaptive use of multiple formats (adaptive preprocessing and format conversion) and the use of forward analysis. The GA supports the transmission and display of 1280 H $\times$ 720 V [SMPTE , 1995a, 1995b] active pixels at 60, 30, and 24 progressive frames/s 1920 H $\times$ 1080 V [SMPTE , 1995a, 1995b] at 30 and 24 progressive frames/s, and 1920 H $\times$1080 V at 30 interlaced frames/s. Except for the interlaced format, all high definition formats have square pixels and progressive scan for interoperability with computer display systems and graphics generation systems. In addition, a degree of compatibility with NTSC is retained with support of 59.94, 29.97, and 23.98 frames/s. A minor consequence of MPEG-2 compatibility is that the encoder and decoder must code 1088 lines while the receiver only displays 1080 lines. This avoids subdivision of 16 $\times$ 16 macroblocks with an insignificant impact on picture quality since the extra 8 lines are normally blank.

Samsung Exhibit 1022, Page 171 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

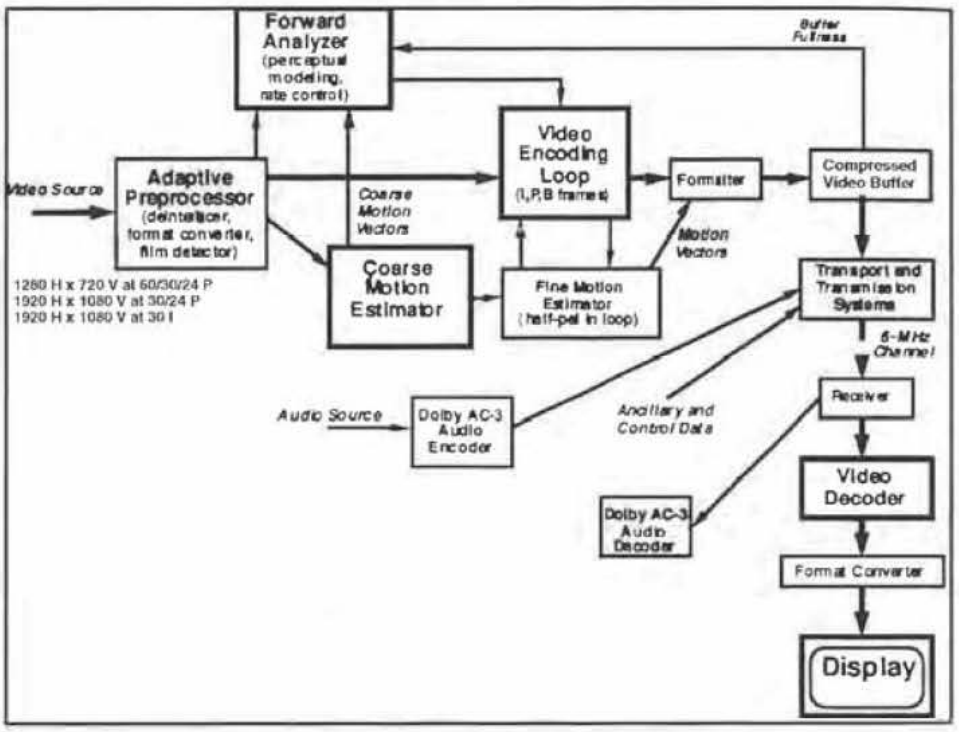


FIGURE 103.1 Grand Alliance system overview.

TABLE 103.1 Standard Resolution Formats Supported by the GA

Image Size	Aspect Ratio	Frame Rate[a] and Int vs Prog, frames/s
1920 H × 1080 V	16 : 9 square pixels	30 Int; 30 and 24 Prog
1280 H × 720 V	16 : 9 square pixels	60, 30 and 24 Prog
704 H × 480 V	4 : 3 or 16 : 9	30 Int; 60, 30 and 24 Prog
640 H × 480 V	4 : 3 square pixels	30 Int; 60, 30 and 24 Prog

[a]Note: All frame rates are optionally reduced by 1000/1001 : ~59.94, ~29.97, ~23.976.

More recently, the GA has included a complete set of standard definition formats which are compatible with VGA computer displays and digital video standards. These formats are included in Table 103.1.

One consequence of incorporating multiple format capability is the need for a scanning format transconversion subsystem before encoding and after decoding. The ability to decouple the transmission or coding format from the source format can result in higher decoded picture quality. For example, if interlaced (Int) video is presented to the system, then the user has the option of deinterlacing followed by transconversion to one of the progressive (Prog) formats (typically, 1280 × 720 at 60 frames/s). In addition, the output of the decoder must be converted to drive a given display format.

103.4 HDTV System Descriptions

Scanning Formats

Psychovisual research [Netravali, 1988] also indicates that loss of spatial resolution in moving areas of a picture is less visible than in stationary areas. Interlaced scanning is an attempt to take advantage

Samsung Exhibit 1022, Page 172 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

of this for compression by spatiotemporal subsampling. Interlaced scanning was attractive for early TV systems (NTSC, PAL, and SECAM) because higher static spatial resolution could be achieved for still picture areas, given the bandwidth limitations of early cameras and displays. However, spatial and temporal aliasing will cause annoying artifacts for certain types of vertical detail (e.g., computer generated images, venetian blinds, and text). In this case, the effective vertical resolution must be greatly lowered by filtering to minimize the appearance of aliasing artifacts. In addition, motion compensated compression efficiency using interlaced scenes is much lower than that for the equivalent progressive scene. Two of the previously proposed HDTV systems (DSC and CCDC) use progressive scanning with 720 active lines per frame (1/59.94 s). The other two systems (AD-HDTV and Digicipher™) use interlaced scanning with 480 lines per field (1/59.94 s) or 960 lines per frame (1/29.97 s).

The choice of video scanning format for an advanced television (ATV) system depends on several competing factors. These factors are not only the subject of technical debate but economic and political debate as well. The GA approach to this issue is to support a reasonable variety of source, transmission, and display formats without adding excessive cost to the decoder. The FCC HDTV standard will primarily specify the transmission formats, whereas the choice of source and display formats will be more subject to forces within the production equipment and display equipment markets respectively.

All of the progressive GA formats have square pixels. One interlaced transmission format is supported mainly because of concerns about the near-term availability of progressive scan HDTV cameras. However, the GA has agreed to work toward the elimination of interlace in the transmission path in the future.

The progressive formats allow a trade between spatial and temporal resolution, and also between resolution and coding artifacts. In the case of 24- or 30-Hz film, the spatial resolution can be 1280×720 or 1920×1080 depending on the scene complexity. If the scene complexity is very high, then a choice of 1280×720 will reduce coding artifacts.

The display format is independent of the transmission formats since the display is not expected to switch between formats instantaneously or at all. A GA compliant decoder will provide frame buffering for all formats, and transconversion will be used to derive the display format. If interlace continues to be used for transmission (as opposed to source video storage or capture), then a deinterlacer must be provided in every receiver with a progressive display.

Motion Compensation

Temporal redundancy is removed by motion estimation and motion compensation [Netravali, 1988]. Motion vectors are computed between frames in the encoder and transmitted to the decoder. The GA system does not perform motion compensation on all frames. The I-frames are coded in isolation without any temporal prediction. The P-frames are coded using forward motion compensation starting with the last I-frame. The B-*frames* are coded using the surrounding I- and/or P-frames to perform forward and backward motion compensation. In addition, a given block in a P- or B-frame may be coded with or without motion compensation. One sequence of I-, B-, and P-frames is known in MPEG-1 and 2 as a group of pictures (GOP).

Motion Estimation

The motion estimator operates only on the luminance images and computes motion vectors using block matching. A motion vector is the spatial displacement between a block of pixels in the current frame and the most similar block in the previous frame (except for B-frames). The similarity or distance measure is the mean of the absolute value of the differences between corresponding pixels in the two blocks. Other distance measures (e.g., mean square error) may provide improved performance but are much more costly to implement than the mean absolute error.

Samsung Exhibit 1022, Page 173 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

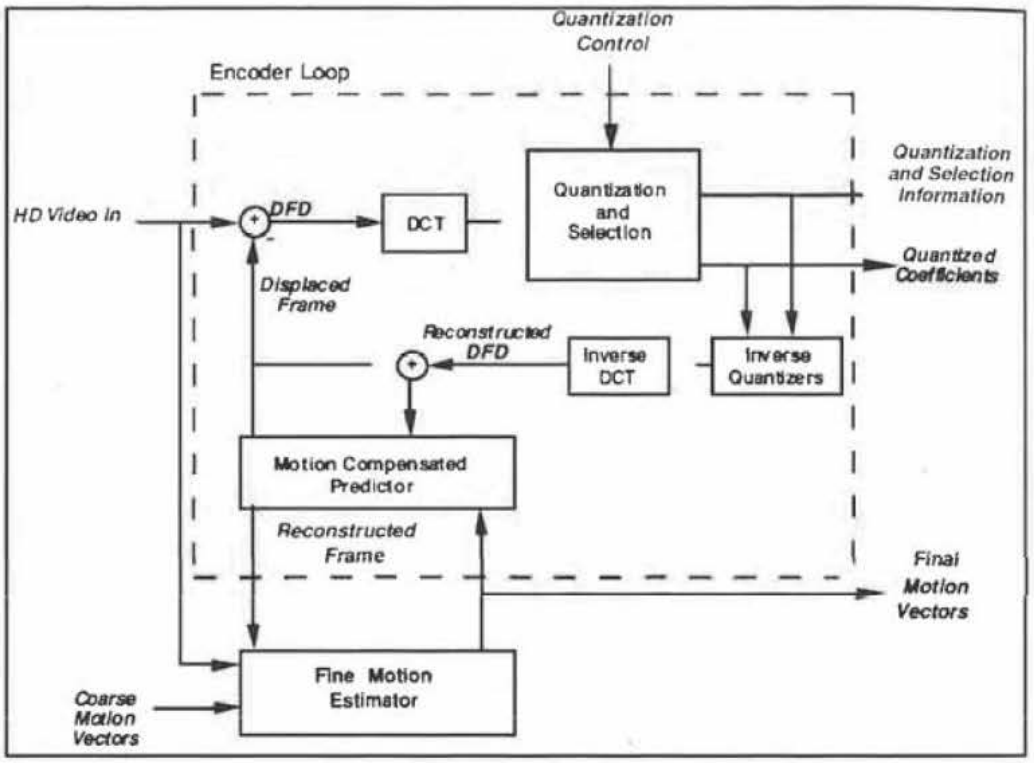


FIGURE 103.2 Video encoder loop.

The number of block matches (and therefore the motion estimator complexity) is proportional to the search area. The GA system uses hierarchical motion estimation to reduce implementation complexity. As shown in Fig. 103.2, the coarse motion vectors, which were computed on the original frames, are used as a reference for the fine and final motion vectors. The reconstructed frame is searched for a best match in the fine motion estimator but for greater hardware economy the search area is much smaller than the coarse stage. In the GA system a total search range of ±64 H and ±32 V is provided by the two stages for B-frames and ±128 H for P-frames. This exceeds the motion estimator range of the predecessor systems.

Predictive Motion Compensation Loop

In the motion compensated prediction loop shown in Figs. 103.1 and 103.2, the motion vectors are applied to the reconstructed frame buffer, which contains a prediction of the decoded picture. The motion compensated predicted frame or displaced frame is subtracted from the original frame to produce the displaced frame difference (DFD). The DFD will be near zero intensity for areas with predictable motion, and nonzero where detail was uncovered or motion was not just a translation in the image plane.

The DFD is the signal which results from the removal of temporal redundancy. An 8×8 **discrete cosine transform (DCT)** is applied to the DFD in preparation for removal of spatial redundancy by coefficient quantization. A reconstructed form of the DFD is then produced by inverse quantization and inverse DCT. The reconstructed or predicted DFD is added to the motion compensated frame to produce the next reconstructed frame, which completes the coding cycle. **Coding artifacts** contained in the reconstructed frame from the current and previous loop cycles can be then be corrected in the DFD coding process.

Samsung Exhibit 1022, Page 174 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

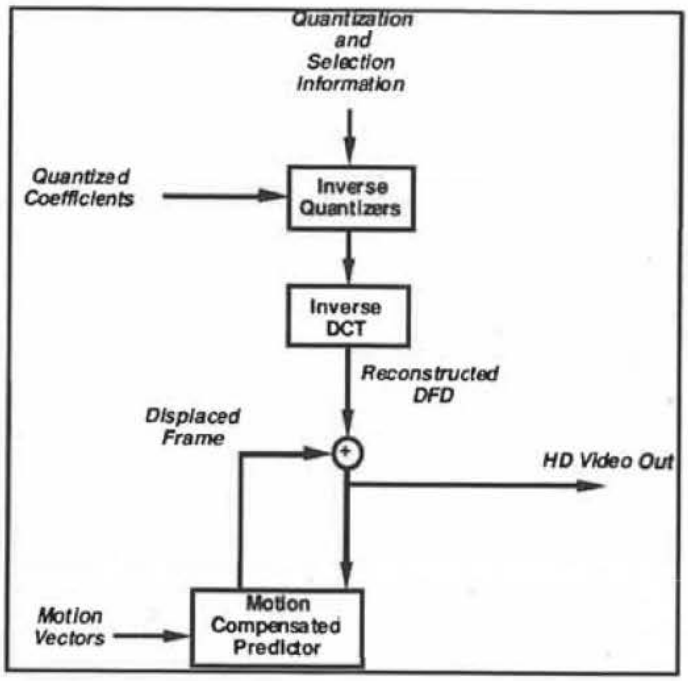


FIGURE 103.3 Video decoder loop.

Decoder Loop

The decoder loop is shown in Fig. 103.3. After deformatting the coded video bit stream, inverse quantization of DCT coefficients and inverse DCT, the reconstructed DFD enters the decoding loop. The decoded motion vectors are applied to the reconstructed frame, resulting in the reconstructed displaced frame. The reconstructed DFD is then added to the reconstructed displaced frame to produce the reconstructed frame.

A GA-compliant decoder must be able to decode bit streams which switch between GA scanning formats. The resulting video must be displayed without loss of picture at the transitions.

Refreshing

The video encoding process exploits temporal redundancy by temporal or inter-frame prediction. This prediction takes the form of motion vectors and an error signal resulting from applying the motion vectors. However, a pure motion compensated prediction loop is not practical without some form of prediction imperfection (i.e., forced *intraframe* coding). If a sequence of input images were perfectly predictable, the decoder would not receive any error signal and therefore could not reconstruct the picture after initialization. Decoder initialization is required after the channel is changed, or the signal is lost and then recovered.

Consider the coding of a still image. The motion vectors and the perfect DFD for a coded still will be zero. The reconstructed DFD in the decoder will also be zero. If the decoder were started with a sequence of zero DFD frames, then a blank or zero reconstructed frame would result. If the viewer changed to a channel transmitting a coded still, the still would therefore never appear. One way to prevent this occurrence is for a portion of the original picture to be mixed with the DFD in the encoder to allow the decoder to synchronize to the encoder after decoder initialization. This also improves recovery from channel impairments. The speed of recovery from channel changes and channel impairments is directly proportional to the amount of original in the mix. Both intraframe

Samsung Exhibit 1022, Page 175 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

coded blocks (I-blocks) and frames (I-frames) are used to provide a mixture of predicted and original video in the transmitted bit stream. The use of I-blocks in a regular pattern that covers the entire image over multiple frames is known as progressive refreshing.

The space between I-frames and the number of consecutive B-frames is variable in the MPEG and the GA systems. Intraframe coding can be done using either I-frames or periodic I-blocks. Since intraframe coding produces a much higher bit rate than interframe coding, I-block refreshing provides easier control of the buffer fullness. However, if one desires to insert a commercial or other compressed video directly into a compressed program, I-frames provide clean insertion points.

Adaptive Selection and Quantization of Coefficients

Sufficient compression cannot be achieved unless a large fraction of the DCT coefficients are dropped and therefore not selected for **quantization** and transmission. The coefficients that are not selected are assumed to have zero value in the decoder. Several different schemes are used in the digital HDTV systems to code the coefficient selection information.

The GA system encodes the selections and runs of zeroes following a zigzag pattern through the array of frequency-ordered coefficients. The DC coefficients are coded differentially to take advantage of high-spatial correlation. The GA system uses one uniform quantizer for coefficient quantization. This method of coding the coefficient selection information and DC coefficients is used in MPEG-1 and 2.

Forward Analyzer

The forward analyzer manages the compressed video buffer and the distribution of quantization error in each coded picture. This is accomplished by monitoring the fullness of the buffer and controlling the compressed video data rate by adapting the amount of quantization error in a visibly uniform manner. The ability of the forward analyzer to anticipate changes in scene complexity is enhanced by the use of the coarse motion vectors which are computed before the picture reaches the encoding loop.

The coefficient quantization process reduces the number of levels used to represent a particular coefficient. This process is controlled by the forward analyzer, which inputs the fullness of the compressed video buffer, coarse motion vectors, and the original pictures, as shown in Fig. 103.1. The coarse motion vectors are used to form an ideal DFD, which provides a prediction of scene complexity ahead of the encoder loop. This prediction reduces the chance of buffer overflow or visible distortion by sudden increases in coarse quantization.

Variable Length Coding

The data representing information such as motion vectors, quantizer selection patterns, and transform coefficients are seldom statistically uniform. Usually, the data are statistically clustered, and the probability distributions can be estimated from analysis of real scenes. The use of variable length codes (VLCs) [Huffman, 1952] takes advantage of this statistical nonuniformity by assigning short codewords to the most frequent values, and assigning longer words to less frequent values. The GA system also provides separate VLC tables for inter- and intracoded coefficients and intra-DC coefficients as specified in the MPEG-2 standard.

The VLC decoder detects a unique start code, which indicates the beginning of a series of VLCs. The 8×8 blocks of pixels are grouped into 16×16 macroblocks. Horizontal runs of macroblocks are grouped into slices where each slice starts with a start code. Short slices reduce error propagation but generate more coding overhead. The GA constrains slices to start at the beginning of a macroblock row.

Samsung Exhibit 1022, Page 176 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

Buffer Control

Motion compensation, adaptive quantization, and variable length coding produce highly variable amounts of compressed video data as a function of time. For example, the compressed bit rate after a scene change can be several times greater than the bit rate in the channel. Therefore, compressed video data buffering in the encoder and decoder is required for efficient channel utilization.

Buffer size is constrained by the maximum tolerable delay through the system and by cost. The fullness of the buffer is controlled by adjusting the amount of distortion or quantization error in each image. In the encoder, the buffer will fill more quickly if the distortion is low. A feedback control system is required to regulate the distortion level, which controls the buffer fullness to prevent overflow. The design of this control system is complicated by the following:

- Delay between a change in the distortion level and the subsequent change in buffer fullness
- Perceptual constraints on the instantaneous and average distortion levels
- Difficulties in modeling the bit rate as a function of distortion level

The visibility of distortion due to an increase in scene complexity is minimized by smoothly increasing the distortion level. This is facilitated by accurately modeling the rate vs distortion level, because an accurate model allows the desired buffer fullness to be achieved for each frame. MPEG-2 specifies a standard rate control model which is parameterized. The parameter bit rate value must not exceed 48,500 for terrestrial broadcast and 97,000 for cable service. The maximum allowable bit rate for a GA decoder must be less than 80 Mb/s. The channel buffer size equals 8 Mb and the buffer size value (vbv) must not exceed 488.

Compressed Video Formats

The compressed video data must be organized into a format that can be reliably transmitted through an imperfect channel and can allow the decoder to recover quickly from the loss of data in the channel. In addition, storage and manipulation of the compressed video and transmission over alternate media (e.g., packet networks) motivates the inclusion of block-, frame-, and program-type descriptors in the format.

Robust transmission is primarily achieved by dividing the compressed video bit stream into fixed length blocks and applying error correction processing to each block. Error correction overhead or parity bits are added to each block before transmission. The receiver is then able to determine if a given block was corrupted in transmission. If a block was corrupted, error concealment is performed in the decoder to minimize the visual impact of the error.

The amount of picture that is degraded by the loss of a block of compressed video data depends on the format of the compressed video bit stream. Since most of the data consists of concatenated variable length codewords, VLC restart pointers must occur at regular intervals in the bit stream. More frequent placement of these pointers will reduce the amount of picture degradation but will also consume more of the bit rate.

GA System Layer

The GA transport is compatible with the MPEG-2 systems layer [ISO, 1994a, 1994b]. As shown in Fig. 103.4, the GA system uses a 188-byte transport packet, which is transmitted to the decoder after trellis coding and Reed–Solomon coding using 8-level vestigial sideband modulation (VSB) [Feher, 1987]. The link header provides simple packet synchronization and identification. A bit also indicates the presence of an adaptation header, which provides clock synchronization, and is also used for video, audio, and data presentation. The choice of a 188-byte packet length is a compromise between the desire to minimize header overhead with longer packets and the need for frequent headers for quick error recovery. Conveniently, each GA transport packet fits into 4 asynchronous transfer mode (ATM) cells with or without the link header.

Samsung Exhibit 1022, Page 177 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

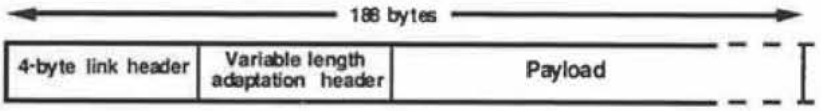


FIGURE 103.4 Grand Alliance transport packet.

Each transport packet header contains a packet identifier (PID) which indicates that audio, video, data, or program map table (PMT) is contained in the packet. The PMT contains the definition of the programs and must be updated at least every 400 ms. The program numbers are associated with the corresponding PMT PIDs via the program association table (PAT). The Advanced Television System Committee (ATSC) has specified additional constraints not detailed here.

GA Transmission System

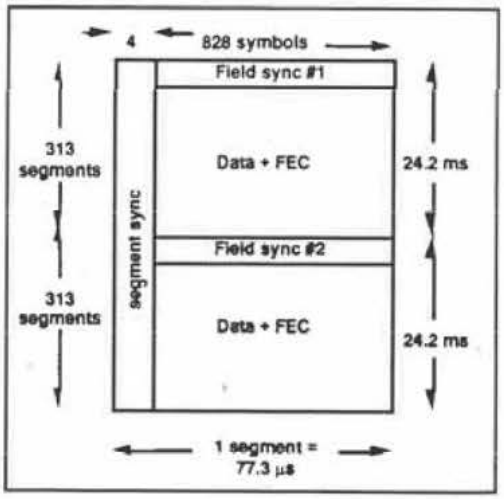


FIGURE 103.5 VSB data frame.

As required by the FCC, a key feature of the GA system is its ability to reliably deliver ~20 Mb/s through the terrestrial TV broadcasting environment using one 6-MHz channel. The FCC recommends that each TV broadcaster be provided with an HDTV channel with equivalent service area, and that the current TV service is not significantly impacted by the HDTV service. These requirements are met by a combination of eight-level digital amplitude modulation (3 b/sample), Reed–Solomon forward error correction (R-S FEC) with 20 R-S parity bytes, a pilot tone, 1/16 field interleaving for burst immunity, 2/3 rate trellis coding, an NTSC rejection filter, and adaptive channel equalization. The suppressed carrier vestigial sideband amplitude modulation of the 8-level samples or symbols (8 VSB) is optimal for the noise- and interference-laden terrestrial broadcast environment, and the VSB signal appears as low-level random noise when interfering with NTSC. The 8 VSB is necessarily coupled with trellis coding for robust transmission. The 8 VSB alone provides 32.3 Mb/s (10.8 megasymbols/s) and the addition of trellis coding drops the payload to 21.5 Mb/s. RS-FEC further reduces the payload to 19.3 Mb/s. Figure 103.5 shows a VSB data frame. The GA also provides an optimal solution for cable TV with 16 VSB. Since trellis coding and its associated bit rate overhead are not needed for cable TV, the GA system can deliver 43 Mb/s over one 6-MHz cable channel. At the present time, the GA system does not support the type of switching between modulation modes (4, 8, or 16 VSB) required to avoid the cliff effect. The ACATS did not strongly support this feature after the first round of tests.

The small pilot tone and NTSC rejection filter specifically target interference from adjacent and frequency colocated NTSC channels. The pilot tone slightly increases the ratio of peak power to average power of the broadcast signal, which slightly increases visibility of the HDTV signal in some NTSC receptions. Its position in the band avoids NTSC carrier peaks primarily from sync signals. NTSC rejection filtering is automatically switched in when needed and slightly reduces the noise immunity of the HDTV reception. Additional synchronization symbols are embedded periodically in the modulated signal, which are not interleaved, RS, or trellis coded.

The performance of the GA-transmission system was tested by ACATS in terms of relative service area compared to the companion NTSC transmitter. The results of laboratory tests using terrestrial RF transmission simulators were input to a computer program which optimized the

Samsung Exhibit 1022, Page 178 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

power, frequency assignment, and location of each GA transmitter. The program output both service areas and viewer populations. The number of stations with less HDTV service area than NTSC is estimated to be about 107 out of 1657, and 187 of the HDTV stations would have at least 20% more service area than NTSC.

103.5 Scanning Format Applications

The GA supports a range of scanning formats for coding and transmission. A given encoder input format does not necessarily indicate the use of that scanning format for encoding and transmission. Table 103.2 shows the advantages of each GA transmission format and indicates how a given encoder input format affects the choice of transmission format. Table 103.3 shows the relationship between transmission formats and display formats.

TABLE 103.2 Applications of Encoder Input Format/Transmission Format Combinations

	Transmission Formats			
Encoder Input Formats	720 × 1280P at 60 frames/s	720 × 1280P at 24/30 frames/s	1080 × 1920I 30 frames/s	1080 × 1920P at 24/30 frames/s
720×1280P at 60 frames/s	Sports, concerts animation, upconverted 480I, commercials, graphics	Adaptively reduce frame rate for lower bit rate or better picture when motion is slow	Not desirable	Adaptively increase spatial resolution and reduce frame rate for high detail/slow motion scenes
720 × 1280P at 24/30 frames/s	Adaptively increase frame rate by frame interpolation	Complex film scenes, complex graphics, complex animation	Not desirable	Adaptively increase spatial resolution
1080 × 1920I at 30 frames/s	Deinterlaced scenes shot with current HDTV cameras	Deinterlaced complex scenes with reduced frame rate for lower bit rate or better picture when motion is slow	Interlaced camera generated scenes/no deinterlacer available	Deinterlaced scenes with reduced frame rate for lower bit rate or better picture when motion is slow
1080 × 1920P at 24/30 frames/s	Adaptively increase frame rate by frame interpolation while reducing spatial resolution for fast motion scenes	Reduced spatial resolution for lower bit rate or better picture	Not desirable	Film

TABLE 103.3 Applications of Transmission Format/Display Format Combinations

	Display Formats		
Transmission Formats	720 × 1280P at 60 frames/s	1080 × 1920I at 30 frames/s	1080 × 1920P at 60 frames/s
720×1280P at 60 frames/s	No conversion required	Have 1080 × 1920I display	Have 1080 × 1920P display
720 × 1280P at 24/30 frames/s	Use 3:2 or 2:2 pulldown	Have 1080 × 1920I display Need 3:2 or 2:2 pulldown	Have 1080 × 1920P display Need 3:2 or 2:2 pulldown
1080 × 1920I at 30 frames/s	Deinterlace in decoder to drive 720 × 1280P display	No conversion required	Deinterlace in decoder to drive 1080 × 1920P display
1080 × 1920P at 24/30 frames/s	Reduce spatial resolution and use 3:2 or 2:2 pulldown to drive 720 × 1280P display	Use 3:2 or 2:2 pulldown	Use 3:2 or 2:2 pulldown

Samsung Exhibit 1022, Page 179 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

HDTV Interoperability with Computers

Personal computers (PCs) today are commonly equipped with high-resolution progressive scan color monitors. Since the most costly part of an HDTV receiver is the display, the early consumers of HDTV may well be multimedia PC owners with HD video decoder cards. For example, the 1280 H × 720 V GA format fits neatly into the widely used 1280 H × 1024 V display with spare screen space for additional windows. High-resolution video on a PC is especially suited for interactive entertainment, which typically requires readable text, and the variety of input devices found on todays PCs. Of course, HDTV on a PC is also ideally suited for education, training, industrial imaging, video and film production, especially when postproduction processing is used. When widescreen HDTV displays become consumer products, those with progressive scanning will also be usable as computer displays.

The compressed HDTV bit stream could be received from terrestrial or satellite broadcast, via coaxial cable (cable TV), or high speed computer network. The bit rate for 24-frame/s film is easily limited to 10 Mb/s (for the 1280 H × 720 V format) for movie services over a local area network. Local storage of compressed HDTV is out of range of current computer disc-read-only memory (CD-ROM) drives but reliable storage and playback using consumer grade video tape technology has been demonstrated.

Migration to 1080 Progressive with 60 Frames Per Second

Enthusiasm has emerged to achieve compression of 1920 × 1080 at 60 frames/s within the 6-MHz terrestrial broadcast channel. This may not be possible for all scenes but may be achievable some of the time. In fact, the deinterlacing process produces an intermediate format of 1920 × 1080 at 60 frames/s which could be used as an input to an augmentation scheme, as shown in Fig. 103.6. A 30 frame/s subset would be coded in the standard way while the remaining frames would be predicted by an interpolator. The difference between the predicted interpolation and the actual frame would be coded and sent to the decoder. Appropriately computed motion vectors could be used for both motion compensation and interpolation. A similar scheme would start with 24- or 30-frame/s film and generate interpolated frames. This scheme would have the advantage of

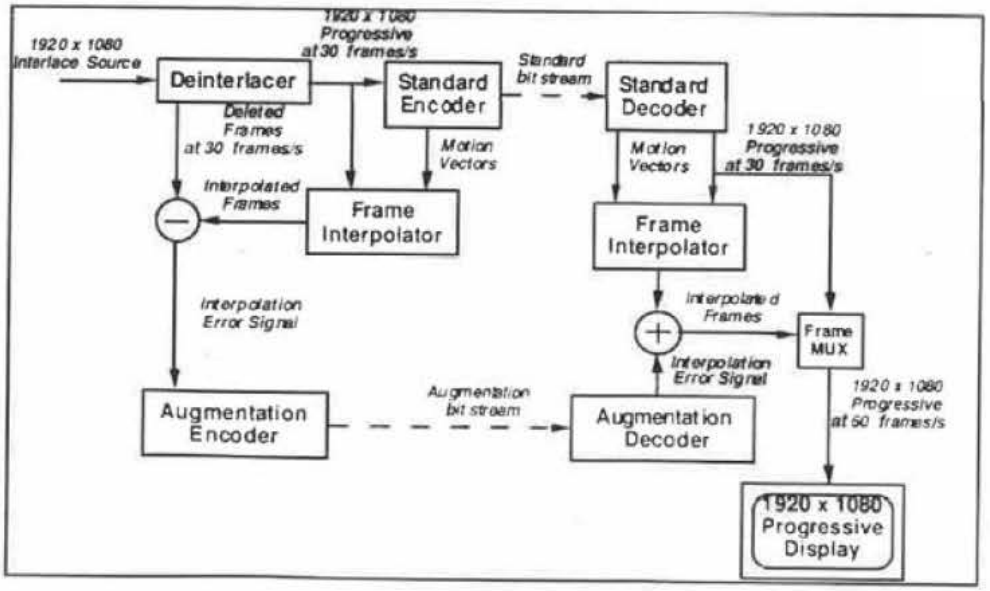


FIGURE 103.6 Possible migration scenario from interlaced source to 1080 progressive display.

Samsung Exhibit 1022, Page 180 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

providing improved temporal performance for existing film material when properly applied. A variety of migration schemes are currently under study for selection by the GA.

Compressed Video Storage and Manipulation

Techniques for efficient storage and manipulation of the compressed video (and audio) data will be necessary to provide consumers with video cassette recorder (VCR) capabilities such as speed search and freeze frame. Since all of the decoders considered here contain reconstructed frame storage, freeze frame is simply provided. Speed search can also be provided by all of these systems with varying degrees of efficiency and picture quality. This function is provided by subsampling the blocks of compressed video data from tapes that are moving at higher than normal speed across the head. The block descriptors are then used to partially reconstruct the video and the picture quality is directly proportional to the number of intraframe coded blocks in the compressed video bit stream. The ability to decode randomly selected portions of compressed video data also facilitates multimedia computing applications, where multiple received and stored video data streams could be manipulated quickly within a window system.

103.6 Conclusions

The GA provides a video compression solution for HDTV which is flexible enough to meet the needs of a wide variety of industries. A rational set of formats with simple conversions between them is provided to accommodate broadcasting, the computer/multimedia industry, computer graphics, industrial imaging, and the national information infrastructure. For example, 1280 × 720 is a convenient subset of a 1280 × 1024 graphics monitor, and the lower frame rate of film allows the compression of 1920 × 1080 for entertainment applications. The use of MPEG-2 syntax provides interoperability with other standards, provides flexible data services at the transport level, and facilitates industry compliance, acceptance, and cost effectiveness for the consumer. Since every GA decoder will accept the variety of GA formats, the future migration of the system to higher quality with evolving encoding methods is ensured.

Defining Terms

Chrominance: Color difference signals formed by weighted sums and differences of red, green, and blue channels.

Coding artifacts: Mostly block boundaries that come and go in detailed or noisy areas of the coded picture.

Discrete cosine transform (DCT): Converts 8 × 8 block of pixels into 8 × 8 block of frequency coefficients, which indicate energy in combined horizontal and vertical frequency spectrum.

Interlace artifacts: Primarily 30-Hz flicker from display of high vertical detail on an interlaced monitor.

Quantization: Mapping a set of intensity values to a smaller set of intensity values resulting in some loss of intensity information and possible coding artifacts.

Taboo channels: The 6-MHz channels, which are unused because of excessive interference from other NTSC channels.

Terrestrial broadcast: Television transmission from a ground station to homes.

References

ATRC. 1992. *Advanced Digital Television System Description,* Jan., Advanced Television Research Consortium.

Samsung Exhibit 1022, Page 181 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540

ATSC. 1995. *ATSC Digital Television Standard,* Document A/53, Sept. 16., Advanced Television Systems Committee.

Feher, K. 1987. *Advanced Digital Communications,* Prentice–Hall, Englewood Cliffs, NJ.

GI. 1991. *DigiCipher™ HDTV System Description,* Aug., General Instrument Corporation.

Huffman, D.A. 1952. A method for the construction of minimum redundancy codes. *Proc. IRE* 40, 1089.

ISO. 1991. ISO CD 11172-2: Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Nov., International Standards Organization.

ISO/IEC. 1994a. MPEG-2 systems. ISO/IEC IS 13818-1, International Standards. Organization/ International Electrochemical Commission.

ISO/IEC. 1994b. MPEG-2 video. ISO/IEC IS 13818-2, International Standards. Organization/International Electrochemical Commission.

LeGall, D. 1991. MPEG: A video compression standard for multimedia applications. *Trans. ACM.*

MIT. 1992. *Channel Compatible Digicipher HDTV System.* American Television Alliance, April, Massachusetts Institute of Technology, Cambridge, MA.

Netravali, H. 1988. *Digital Pictures: Representation and Compression.* Plenum Press, New York.

Service, F.A.C.o.A.T. 1993. Federal Communications Commission advanced television system recommendation. *IEEE Trans. Broadcasting,* 39(1):3–245.

SMPTE. 1995a. Proposed Standard for television, 1280 × 720 scanning and interface. SMPTE S17.392. Society of Motion Picture and Television Engineers.

SMPTE. 1995b. Standard for television, 1920 × 1080 scanning and interface. SMPTE 274M. Society of Motion Picture and Television Engineers.

Further Information

An excellent review of picture processing is found in Netravali [1988]. The ATSC Digital Television Standard (see references) is a well written and concise description of the proposed FCC HDTV standard. For a description of the ATTC test results see the Final Technical Report, FCC ACATS, October 31, 1995. A guide for using the Grand Alliance system can be found in the Guide to the use of the ATSC Digital Television Standard, Doc. A/54, October 4, 1995. A description of the Dolby AC-3 audio compression system is found in Digital Audio Compression Standard (AC-3), Doc. A/52 May 24, 1995.

Samsung Exhibit 1022, Page 182 of 182
Samsung Electronics Co., Ltd. et al. v. One-E-Way, Inc.
IPR2025-01540